UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in the Design of Experiments

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Economics

by

Stefan Firaydun Faridani

Committee in charge:

Professor Graham Elliott, Chair
Professor Xinwei Ma
Professor Gareth Nellis
Professor Paul Niehaus
Professor Kaspar Wuthrich

2024

The Dissertation of Stefan Firaydun Faridani is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To my wife, to my sister, and to my mother and father.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Elliott for his support as the chair of my committee and Professor Niehaus for his patient and tireless mentorship.

Chapter 1, in part is currently being prepared for submission for publication of the material. The dissertation author was the sole author of chapter 1.

Chapter 2, in part is currently being prepared for submission for publication of the material. The dissertation author was one of two primary authors of this material. The other was Professor Paul Niehaus.

Chapter 3, in full, is an unpublished report issued midway through a multi-year field experiment. The dissertation author wrote the full text of this chapter and provided technical work at many stages of the field experiment from October 2021 until the present. The co-authors are Moustafa El-Kashlan, Vesall Nourani, and Sara Restrepo-Tamayo.

VITA

2017	Bachelor of Arts, Macalester College
2017–2018	Research Assistant, Marshall Institute, London School of Economics
2019–2021	Teaching Assistant, Department of Economics University of California San Diego
2021–2023	Research Assistant, University of California San Diego
2024	Doctor of Philosophy, University of California San Diego

FIELDS OF STUDY

Major Field: Economics (Econometrics)

ABSTRACT OF THE DISSERTATION

Essays in the Design of Experiments

by

Stefan Firaydun Faridani

Doctor of Philosophy in Economics

University of California San Diego, 2024

Professor Graham Elliott, Chair

This dissertation presents three essays in the design of randomized experiments. Chapter 1 “Testing for Underpowered Literatures” proposes a novel estimator consistent for the expected number of statistically significant results that a set of experiments would have reported had their sample sizes all been counterfactually increased by a chosen factor. An application to randomized controlled trials published in top economics journals finds that doubling every experiment’s sample size would only increase the power of two-sided t-tests by 7.2 percentage points on average. Chapter 2 “Linear Estimation of Global Average Treatment Effects” studies the problem of estimating the average causal effect of treating every member of a population, as opposed to none, using an experiment

that treats only some. This differs from the usual average treatment effect in the presence of spillovers. We find the fastest possible rate of convergence and provide experimental designs and estimation procedures that achieve this rate. We also propose modifications to the regression-based estimator common in applied work to guarantee its rate-optimal consistency. Chapter 3 “Evaluation of a Teacher Training in Uganda: Specifications for Endline” is a midline report for an ongoing field experiment in Uganda. The experiment evaluates the impact of a general skills teacher training program in rural Uganda. In the trial, we offer the program to a sample of 640 teachers in 39 secondary schools.

Chapter 1

Testing for Underpowered Literatures

1.1 Introduction

A key tradeoff in experimental design is balancing the risk of drawing an erroneous conclusion with project cost. Nearly all experiments published in top economics journals use t-tests to interpret their results. Sampling variance means that the t-test can fail to reject the null hypothesis of zero treatment effect even when there is in fact an effect of meaningful magnitude. Every experiment therefore runs the risk of a false negative. Increasing the sample size raises statistical power and reduces the probability of a false negative but also requires additional funding. A central question faced by funders and researchers is whether to concentrate resources into a small number of experiments or to spread them across many.

This paper provides meta-analysts and research funding organizations with a statistical procedure to estimate how much larger the expected fraction of statistically significant t-scores would have been had every experiment in some collection been counterfactually run with c^2 times the sample size where $c > 1$. This power gain is a number between zero and one that I call Δ_c. If this quantity is large within a given scientific literature or funding initiative, then the rejection decisions of t-tests reported by that population of experiments are sensitive on average to sample size choices. In this case, the returns to concentrating resources into fewer and larger experiments could be

relatively high.

To estimate Δ_c with the proposed method, the meta-analyst needs a dataset of n t-scores reported by a collection of experiments. The statistical procedure works by first finding the distribution of true intervention treatment effects that best fits a smoothed version of the empirical distribution of t-scores. The fitted distribution of true effects is then integrated to calculate a point estimate of Δ_c. I show that this procedure is consistent, asymptotically normal, and converges in a power of n.

The main contribution of this paper is that it removes the reliance on functional form assumptions made by related meta-analyses that study power. The methods of Ioannidies et al. (2017), Brunner and Schimmack (2020), and others are only consistent when the population distribution of true intervention treatment effects has a specific shape. Functional form restrictions of this kind need not hold in practice. In contrast, the only assumption made by this paper on the distribution of true treatment effects is that it has a probability density function with bounded height.

A second contribution of this paper is to make its method robust to simple forms of publication bias. Recent empirical evidence shows that statistically insignificant t-scores are less likely to be published in academic journals (Franco et al., 2014; Andrews and Kasy, 2019). Selective reporting can distort the distribution of reported t-scores and potentially confound estimation. This paper addresses selective reporting by parameterizing a simple model of publication bias, estimating the model, and then reweighting the observed t-scores to remove publication bias. Accommodating this first stage requires that I use a particular smoothing method which results in a different estimator than those used to solve mathematically related *deconvolution* problems like Carrasco and Florens (2011).

An empirical estimate of Δ_c is useful to meta-analysts and funders because they may otherwise lack a way to evaluate how efficiently sample sizes are being chosen in practice. Sample size decisions are typically made under high uncertainty. If funders and

researchers knew ex ante how quickly the statistical power of a proposed experiment responded to its sample size, they could optimally balance power and cost. But power also depends on how effective the treatment under study actually is, a quantity which is ex ante unknown. Grant-giving organizations try to address this challenge by requiring experimenters to collect samples large enough to guarantee at least 80% power to detect a true effect larger than some chosen threshold (Gupta and Kopper, 2023). The aim of these *power calculations* is to direct resources to projects that will only fail to detect an effect when that effect is in fact small.

Power calculations may fall short of this goal in practice because they involve a great deal of guesswork. Statistical power depends on many unknown parameters besides the true effectiveness of the intervention, e.g. the degree of heteroskedasticity of the outcome variable or how heterogeneous treatment effects are across individuals. If the power calculations are inaccurate, then the choice of sample size made on their basis may trade off power and cost sub-optimally.

Even ex post it is difficult to determine what the consequences of a larger sample size would have been for a single experiment. This problem arises because it is not in general possible to infer how powerful an individual experiment was. Plugging the estimated treatment effect into the power function is known to be highly misleading (Hoenig and Heisey, 2001). Experimenters and funders therefore lack a rigorous way to assess how efficiently sample sizes are being chosen in practice. The method proposed in this paper aims to address this need.

This paper applies its method to an empirical question of broad interest: how sensitive are randomized controlled trials (RCTs) published in top economics journals to counterfactual sample size increases? I use the data from Brodeur et al. (2020) which contains t-tests of main hypotheses reported by articles published in 25 top economics journals during 2015-2018. I estimate that counterfactually doubling the sample sizes of every RCT in that set would only increase the expected number of t-scores clearing the

critical value of 1.96 by 7.2 percentage points with a standard error of 2.5.

This power gain is small in comparison to several benchmarks. First, I estimate that doubling the sample size of all non-RCTs in the same dataset would increase average power by 17.3 percentage points which is significantly larger. As a second benchmark, consider a hypothetical literature where every experiment is adequately powered to detect the true effect. By conventional standards, a literature where every experiment were powered at exactly 80% power would meet this criterion (Gupta and Kopper, 2023). For such a hypothetical literature, we can calculate that doubling every sample size would increase power by 17.8 percentage points, which is much larger than the empirical estimate.

I construct a third benchmark using data from the Many Labs systematic replication project (Klein et al., 2014). In this project, 36 laboratories each independently attempted to replicate 11 published effects from laboratory psychology. We would expect these replication experiments to be very well powered because they were designed using previous experimental data and the consequences of failure to detect an already published effect could be great. Yet I cannot reject the null hypothesis that Δ_c is the same for both RCTs in economics and the Many Labs replications. This means that RCTs are on average about as sensitive to sample size as replication experiments that were designed using a great deal of prior knowledge. The non-rejection is not simply from high uncertainty because we can indeed reject the null that Δ_c is the same for non-RCTs in economics vs replications from Many Labs. I also leverage the fact that the Many Labs experiments are replications to estimate power gain conditional on sample true effects. The conditional Many Labs power gain is precisely estimated at 7.8 percentage points with a standard error of 0.6 percentage points and is also statistically indistinguishable from the Δ_c for economics RCTs.

I conclude from this analysis that randomized trials in economics are on average relatively insensitive to counterfactual sample size increases. The policy implication

is that funders and researchers in economics should broadly consider running more randomized trials, not fewer, larger ones.

This paper is organized as follows. Section 1.2 discusses the contributions of this paper to existing applied and theoretical literatures. Section 1.3 sets up the estimation problem without publication bias. Section 1.4 shows identification, proposes an estimator, and derives its rate of consistency. Section 1.5 introduces publication bias. Section 1.6 shows asymptotic normality and discusses inference. Section 1.7 uses simulations to recommend tuning parameters. Section 1.8 presents two empirical applications and Section 1.9 concludes.

1.2 Related Literature

This paper contributes to two literatures. The first is an applied literature that empirically studies statistical power under strict assumptions. Ioannidies et al. (2017) estimates median power in empirical economics under the assumption that papers can be sorted into groups within which every study is estimating the same true effect. A similar assumption is used by DellaVigna and Linos (2022), Arel-Bundock et al. (2022), and Ferraro and Shukla (2023). Other methods instead assume that the distribution of true effects has a specific shape (Brunner and Schimmack, 2020; Sotola, 2023; Lang, 2023). Assumptions like these need not hold in practice and are difficult to check. This paper contributes to this literature by assuming only that the population distribution of true effects has a probability density function with bounded height.

The second related literature is technical. The estimation problem studied in this paper belongs to a class of problems called *deconvolutions* which aim to "de-blur" a smooth probability density. The theoretical econometrics literature has proposed a number of solution methods for deconvolutions (Fan, 1991; Carrasco et al., 2007; Carrasco and Florens, 2011; Racine et al., 2014). But in the presence of publication bias

none of these methods will be consistent. There is growing evidence that statistically insignificant t-scores are less likely to be reported in economics publications (Andrews and Kasy, 2019; Elliott et al., 2022; Brodeur et al., 2020, 2016). Such omissions can create a discontinuity in the probability density of t-scores at the significance threshold (Gerber and Malhotra, 2008; Kudrin, 2023).

Addressing this problem is not as simple as concatenating some existing deconvolution method with publication bias removal. This paper proposes a new deconvolution step in order to manage the interactions between the two stages. The method begins with the same singular value decomposition as Carrasco and Florens (2011) but uses a different kind of regularization called *spectral cutoff* that prevents uncertainty about the extent of selective reporting from magnifying the regularization bias. This choice of smoothing method results in a new estimator that converges at a different rate.[1]

1.3 Setup

First define a key piece of notation. Let $\varphi(z)$ denote the probability density function of the standard normal distribution. Adding a subscript $\varphi_{\sigma^2}(z)$ denotes the density of the normal distribution with variance σ^2. The absence of the subscript means that the variance is unity.

1.3.1 Estimand

Consider a population of experiments. Each experiment studies a unique intervention with its own treatment effect $b \in \mathbb{R}$. The treatment effect b is unobserved, but the experimenter estimates it with an estimator $\hat{b}$ that is unbiased and normally distributed. This means that: $\hat{b} \,|\, b \sim N\left(b, \sigma^2\right)$. The experimenter knows the standard error σ and summarizes the evidence against the null hypothesis of zero treatment

[1] Spectral cutoff regularization is used by Florens et al. (2017) in related but distinct setting. They estimate an inner product, assume more smoothness than we do, and equip their domain and codomain with the same inner product to construct eigenfunctions.

effect by reporting the t-score: $T = \frac{\hat{b}}{\sigma}$. Defining $h \equiv \frac{b}{\sigma}$, makes it possible to express the conditional distribution of T in terms of h only. I call h the "true effect." The ratio T is conditionally normally distributed centered on h with variance 1, i.e. $T \mid h \sim N(h,1)$. Its conditional probability density function $f_T(t \mid h)$ is the following:

$$f_T(t \mid h) = \varphi(t-h) \tag{1.1}$$

T is interpreted as the test statistic of a two-sided t-test of the null hypothesis that $h = 0$. The power of a size-α test run by an individual experiment is the conditional probability that $|T|$ exceeds the critical value $cv(\alpha)$. I suppress the dependence of the critical value on α for ease of notation. I call this conditional probability the *conditional power*. Conditional power can be written in terms of an integral over the normal density.

$$\underbrace{\Pr(|T| > cv \mid h)}_{\text{Conditional Power}} = 1 - \int_{-cv}^{cv} \varphi(t-h)dt \tag{1.2}$$

Conditional power is always unknown. Since h is unobserved, the meta-analyst never gets to condition on it. In practice this means that it is not possible to recover the power of any *individual* experiment using only the t-score that it reported (Hoenig and Heisey, 2001). To see why, notice that Jensen's Inequality guarantees that even though $E[T \mid h] = h$, nevertheless $E[\varphi(t-T) \mid h] \neq \varphi(t-h)$. Invoking consistency and the continuous mapping theorem does not help since the limit of conditional power as the experimental sample size goes to infinity is either α or 1.

The meta-analyst does not need to know the true power of any individual experiment. Instead the meta-analyst wishes to know the expected statistical power of an experiment *randomly drawn* from a population of experiments. I call this expectation the "unconditional power." Unconditional power depends on the distribution of h. Assumption 1, states that the population distribution of h is continuous with probability density

function π_0. This density must have finite height. Assumption 1 is not as restrictive as it seems because the distribution of t-scores under a discrete π_0 can be arbitrarily well approximated by the distribution of t-scores under a continuous π_0.

Assumption 1. *The distribution of h is continuous with bounded probability density function $\pi_0(h)$.*

Unconditional power is defined as the expectation of conditional power over the distribution of h. By Fubini's Theorem, the order of integration can be exchanged and unconditional power can be expressed in the following way.

$$\underbrace{\Pr(|T| > cv)}_{\text{Unconditional Power}} = 1 - \int_{-cv}^{cv} \int_{-\infty}^{\infty} \varphi(t-h)\pi_0(h)dh\,dt \tag{1.3}$$

Experimenters have some control over the power of their experiments even though they do not know h because they can choose the sample size of the experiment. Increasing the sample size of an experiment will increase its power whenever $h \neq 0$. The meta-analyst wishes to learn how much larger unconditional power would have been had the sample sizes of every experiment been counterfactually increased while holding the distribution of true effects constant.

Define the random variable T_c as a t-score randomly drawn from a counterfactual population where every experiment has been run at c^2 times the actual sample size while holding π_0 constant. Since it is counterfactual, no draw of T_c is ever actually observed. Multiplying the sample size by c^2 shrinks the standard error and grows h by a factor of c. Conditional on the true effect, T_c is normally distributed but with a larger mean, i.e. $T_c \,|\, h \sim N(ch, 1)$. The unconditional counterfactual power is defined as the probability that T_c exceeds the critical value. This can be expressed as the following double integral.

$$\underbrace{\Pr(|T_c| > cv)}_{\text{Unconditional Counterfactual Power}} = 1 - \int_{-cv}^{cv} \int_{-\infty}^{\infty} \varphi(t-ch)\pi_0(h)dh\,dt \tag{1.4}$$

The aim of this paper is to determine whether the unconditional power of a given population of experiments is sensitive to sample size increases. This represents an "opportunity missed" because there are in expectation many false negatives that could have been avoided had the experiments been somewhat larger. The next step is to quantify this idea. Define the estimand Δ_c as the power gain resulting from increasing every sample size in the population of experiments by a factor of c^2.

$$\Delta_c \equiv \Pr(|T_c| > cv) - \Pr(|T| > cv) \tag{1.5}$$

Δ_c will be the estimand throughout this paper. Example 1.1 illustrates why I interpret a small value of Δ_c as an indicator of an underpowered literature. The subsequent remarks build further intuition for the estimand.

Example 1. The difference between status quo and counterfactual power depends on the distribution of true effects π_0. When a true effect is extremely small or large, power conditional on that effect is unresponsive to sample size. So literatures that contain many extreme true effects will be less responsive than those with few extreme true effects. The following example illustrates. Consider two different literatures each with their own distribution of true effects. For both literatures, the standard error and sample size are the same with $\sigma = 1$ and $\sqrt{n} = 100$. So for both literatures $h = 10b$. In literature 1, the intervention impact is distributed $b \sim N(0.17, 0.1)$. In literature 2, the intervention impact b is a 50-50 mixture of two normals, one centered on 0.02 and the other centered on 2, both with standard deviation 0.1. Figure 1.1 shows how unconditional power (y-axis) responds to counterfactual sample size increases (x-axis). Increasing n leads to rapid gains for literature 1 but not literature 2 because literature 2 is composed almost entirely of extremely large or small true effects. The task of this paper is to use a sample of t-scores to determine whether they were drawn from a population more like literature 1 or literature 2.

Remark 1. The estimand Δ_c considers a counterfactual where the sample sizes increase but the distribution of true effects π_0 remains constant. This framework is flexible but does have limitations. For example, if increasing the sample size of a field experiment would entail increasing spillover effects then π_0 is different under that counterfactual and Δ_c would not be the estimand of interest.

Remark 2. To build intuition it is helpful to study the possible range of values that the estimand can take. Δ_c is always non-negative because power is non-decreasing in sample size. Δ_c can be arbitrarily close to zero because π_0 can place arbitrarily large probability mass near zero. Finally, Δ_c is maximized for each c when π_0 places all of its probability mass near cv. The maximum possible value that Δ_c can take on depends on c and cv. Figure 1.2 illustrates how the maximum possible value of Δ_c varies with c when we hold cv fixed at 1.96. The figure shows for example that doubling of the sample size of every experiment in a population increases the power gain of a two-sided 5% t-test by at most 0.292.

1.3.2 Convolution Operators

It is possible to think of the random variable T as the sum of h plus an independent "noise" random variable that has the normal distribution, i.e. $T = h + Z$. The sum of two independent random variables is called a *convolution* and an operator that maps the distribution of h into the distribution of h plus an noise is called a *convolution operator*. It will be useful to express the distributions of T and T_c in terms of convolution operators. The first step is to write down the densities of T and T_c as integrals over the distribution of true effects.

$$f_T(t) = \int_{-\infty}^{\infty} \varphi(t-h)\pi_0(h)dh$$
$$f_{T_c}(t) = \int_{-\infty}^{\infty} \varphi(t-ch)\pi_0(h)dh$$

Both densities are the outcomes of closely related mappings of π_0. This type of mapping can be generalized. Consider the operator K_{σ^2} below that maps the density of h to the density of $h+\sigma Z$ where Z is a standard normal random variable independent of h where $\sigma > 0$. The domain of this operator will be defined later on. For now it is enough to point out that K_{σ^2} maps any probability density into another probability density.

$$(K_{\sigma^2}\pi)[t] \equiv \int_{-\infty}^{\infty} \varphi_{\sigma^2}(t-h)\pi(h)dh$$

A key fact about normal convolutions is that they can be decomposed. Lemma 1 says that adding normal noise of variance 1 is equivalent to adding normal noise of variance c^{-2} and then adding further independent normal noise of variance $1-c^{-2}$.

Lemma 1. *For any function π in the domain of $K_{c^{-2}}$, and $c > 0$:*

$$K_1\pi = K_{1-c^{-2}}K_{c^{-2}}\pi$$

Proof: Section 1.13.1.

This decomposition will be useful. It is immediate to see that $f_T = K_{1-c^{-2}}K_{c^{-2}}\pi_0$. Lemma 2 expresses the estimand Δ_c in terms of $K_{c^{-2}}\pi_0$ as well. The motivation for this decomposition is that $K_{c^{-2}}\pi_0$ turns out to be much easier to estimate than π_0 itself because it has already been smoothed out. The next subsection sets up the theory to show why this is the case.

Lemma 2. *If Assumption 1 holds, then:*

$$\Delta_c = \int_{-cv}^{cv} (K_{1-c^{-2}}K_{c^{-2}}\pi_0)[t]dt - \int_{-cv/c}^{cv/c} (K_{c^{-2}}\pi_0)[t]dt$$

Proof: Section 1.13.2.

1.3.3 Singular Value Decomposition

The meta-analyst observes draws from f_T and wishes to estimate Δ_c. To do this it is sufficient to estimate π_0. The problem of recovering π_0 from f_T is severely ill-posed because very large changes in π_0 can result in very small changes in f_T. The intuition is that applying a convolution operator K produces smoothed-out version of the argument. Since π_0 is not necessarily smooth, many of its "high-frequency" features are destroyed by convolution and are therefore difficult to recover from f_T. However, the problem of recovering $K_{c^{-2}}\pi_0$ from f_T is much better-posed. This target density doesn't have many high frequency features because $K_{c^{-2}}$ has already destroyed them.

This intuition can be formalized using the singular value decomposition. Intuitively, the singular value decomposition expresses $K_{c^{-2}}\pi_0$ as an infinite weighted sum of components of π_0 where the weights are guaranteed to decay geometrically fast. This decomposition reveals which features of π_0 are preserved by $K_{c^{-2}}$ and which cannot be recovered. The singular value decompositions presented in this section are modified versions of the decompositions in Carrasco and Florens (2011) and Wand and Jones (1995).

The first task is to precisely define the domain and range of the two convolution operators $K_{c^{-2}}$ and $K_{1-c^{-2}}$. The meta-analyst must start by choosing the scalar $\sigma_Y^2 > 0$. This could be considered a tuning parameter but in this paper I always choose $\sigma_Y^2 = 1$ and never deviate from this choice. Define the following three Hilbert spaces of functions: $\mathscr{L}_Y, \mathscr{L}_X, \mathscr{L}_W$.

$$\mathscr{L}_Y \equiv \left\{ \phi(x) \text{ such that } \int_{-\infty}^{\infty} \phi(x)^2 \varphi_{\sigma_Y^2}(x)dx < \infty \right\}$$
$$\mathscr{L}_X \equiv \left\{ \phi(x) \text{ such that } \int_{-\infty}^{\infty} \phi(x)^2 \varphi_{\sigma_Y^2+1-c^{-2}}(x)dx < \infty \right\}$$
$$\mathscr{L}_W \equiv \left\{ \phi(x) \text{ such that } \int_{-\infty}^{\infty} \phi(x)^2 \varphi_{\sigma_Y^2+1}(x)dx < \infty \right\}$$

These spaces are all very large. Each contains every function in $\mathscr{L}^2(\mathbb{R})$ and every probability density function of a real-valued random variable. Moreover, $\mathscr{L}_Y \subset \mathscr{L}_X \subset \mathscr{L}_W$. Equip each space with the following integral inner products respectively:

$$\begin{aligned}
\langle \phi_1, \phi_2 \rangle_Y &\equiv \int_{-\infty}^{\infty} \phi_1(x)\phi_2(x)\varphi_{\sigma_Y^2}(x)dx \\
\langle \phi_1, \phi_2 \rangle_X &\equiv \int_{-\infty}^{\infty} \phi_1(x)\phi_2(x)\varphi_{\sigma_Y^2+1-c^{-2}}(x)dx \\
\langle \phi_1, \phi_2 \rangle_W &\equiv \int_{-\infty}^{\infty} \phi_1(x)\phi_2(x)\varphi_{\sigma_Y^2+1}(x)dx
\end{aligned}$$

These inner products induce norms. So $||\phi||_Y^2 \equiv \langle \phi, \phi \rangle_Y$. Now the convolution operators can be fully defined by specifying their domains and the spaces that they map into.

$$\begin{aligned}
K_{c^{-2}} &: \mathscr{L}_W \to \mathscr{L}_X \\
K_{1-c^{-2}} &: \mathscr{L}_X \to \mathscr{L}_Y
\end{aligned}$$

Both $K_{c^{-2}}, K_{1-c^{-2}}$ are compact linear operators. This means that they have singular value decompositions. In this case the singular value decomposition will express the outcome of a convolution as a weighted sum of known orthonormal polynomials where the weights are known to decay at an geometric rate. These decompositions are expressed below.

$$\begin{aligned}
K_{c^{-2}}\pi &= \sum_{j=0}^{\infty} \underbrace{\eta_j}_{\text{scalar}} \langle \chi_j, \pi \rangle_W \quad \underbrace{\phi_j}_{\text{polynomial}} \\
K_{1-c^{-2}}g &= \sum_{j=0}^{\infty} \underbrace{\lambda_j}_{\text{scalar}} \langle \phi_j, g \rangle_X \quad \underbrace{\psi_j}_{\text{polynomial}}
\end{aligned}$$

The singular values η_j, λ_j are the sequences of scalars defined below. These decay geometrically fast to zero. The fast rate of decay implies that the problem of recovering

π_0 from $K_1\pi_0$ is ill-posed because components of π_0 with large j play a very small role in $K_1\pi_0$.

$$\eta_j = \left(\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}\right)^{j/2}$$
$$\lambda_j = \left(\frac{\sigma_Y^2}{\sigma_Y^2+1-c^{-2}}\right)^{j/2}$$

The singular functions χ_j, ψ_j, ϕ_i are the generalized probabilist's Hermite Polynomials. Since the expressions for the polynomials are very long, they are relegated to Appendix 1.12.1. There are four important properties of the Hermite polynomials that the reader should take note of. First, the polynomials are normalized so that, for example $\langle \chi_j, \chi_j \rangle_W = 1$. Second, each set forms a complete basis for its corresponding Hilbert space, all of which are weighted $\mathscr{L}^2$ spaces (Johnston, 2014). This means that for any two probability densities π_1, π_2, if $\langle \pi_1 - \pi_2, \chi_j \rangle_X = 0$ for all j, then $\pi_1 = \pi_2$ almost everywhere. Third, the polynomials are orthogonal, so $\langle \chi_j, \chi_k \rangle_W = 0$ when $j \neq k$. Fourth, while these polynomials are themselves unbounded, they are uniformly bounded over all t, j when multiplied by the kernels from their respective inner products (Indritz, 1961). Formally this means that: $\sup_{t,j} \left| \varphi_{\sigma_Y^2+1}(t)\chi_j(t) \right| < \infty$.

Equation 1.6 below expresses the density of T as a linear combination of the Hermite polynomial. Notice that the singular values λ_j and η_j are forcing higher order polynomials to play a small role. The implication for the meta-analyst is that high-frequency information about π_0 is not easy to recover from f_T.

$$f_T = \sum_{j=0}^{\infty} \underbrace{\lambda_j \eta_j}_{\text{scalars}} \langle \chi_j, \pi_0 \rangle_W \underbrace{\psi_j}_{\text{polynomials}} \tag{1.6}$$

Lemma 3 expresses the target estimand Δ_c in terms of the singular values and Hermite polynomials. This gives insight into the ill-posedness. The sequence of coeffi-

cients $\langle \chi_j, \pi_0 \rangle_W$ is sufficient for both the distribution of the data and for the estimand Δ_c. Equation 1.6 shows that the larger j is, the more difficult it is to recover $\langle \chi_j, \pi_0 \rangle_W$ from f_T since it is damped away by the geometrically decaying $\eta_j \lambda_j$. But on the other hand, Lemma 3 shows that the larger j is, the smaller the weight that $\langle \chi_j, \pi_0 \rangle_W$ is given in the estimand because η_j is decaying as well. This illustrates why Δ_c is fundamentally easier to recover from f_T than π_0.

Lemma 3.

$$\Delta_c = \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_0 \rangle_W \left(\lambda_j \int_{-cv}^{cv} \phi_j(t) dt - \int_{-cv/c}^{cv/c} \phi_j(t) dt \right)$$

Proof: Section 1.13.3.

1.4 Identification and Estimation without Publication Bias

This section shows that π_0 is identified and proposes a consistent estimator of Δ_c under the assumption that there is no publication bias. Econometric deconvolution problems are well studied (Fan, 1991; Carrasco et al., 2007; Carrasco and Florens, 2011; Racine et al., 2014). My estimator has an unusual motivation in that its real goal is to recover Δ_c and it must be set up in order to be compatible with an adjustment for publication bias later on. This motivates my choice of the spectral cutoff regularization method. Spectral cutoff has two advantages in this context. First, it eliminates dependence of the smoothing parameter on c. Second, when publication bias is introduced in the next section, spectral cutoff will allow me to concentrate the objective function and estimate publication bias and π_0 one at a time. But to avoid introducing too many ideas at once, I first show how my deconvolution method works when every t-score is reported.

1.4.1 Identification

I prove that π_0 is identified in the standard way by showing that it is the unique minimizer of a population objective function (Newey and McFadden, 1994). I specify the

objective function $Q_0(\pi)$ to be the integral of the squared difference between the true density f_T and the density of T implied by the argument distribution π. The difference is weighted by the normal density. Intuitively, minimizing this objective function is the continuous analogue of a weighted least squares regression.

$$Q_0(\pi) \equiv \int_{-\infty}^{\infty} \left(f_T(t) - \left(\int_{-\infty}^{\infty} \varphi(t-h)\pi(h)dh \right) \right)^2 \varphi_{\sigma_Y^2}(t)dt \tag{1.7}$$

By inspection the objective is non-negative and $Q_0(\pi_0) = 0$. So π_0 is a minimizer. There are several ways to show that this minimum is unique. The following paragraphs provide an argument using the singular value decomposition.[2] This argument is analogous to checking whether a regression matrix in a least squares problem is full rank, which can also be done using the matrix singular value decomposition.

Since the integrand is weighted by the normal density, the objective function is actually the inner product of the difference in densities with itself. This is the norm induced by the inner product $\langle \cdot, \cdot \rangle_Y$. The expression below makes the analogy to regression even more plain.

$$Q_0(\pi) = ||f_T - K_1\pi||_Y^2$$

This new expression says that π_0 is the unique minimizer of Q_0 if and only if convolution cannot map two different densities into the same density. The singular value decomposition reveals why this is the case. Just like in Euclidean space, norms induced by inner products can be expressed as squared sums over orthonormal basis vectors. The singular functions ψ_j form a complete orthonormal basis for $\mathscr{L}_Y$. The population objective function can now be written as the sum of squared differences in basis coefficients.

$$Q_0(\pi) = \sum_{j=0}^{\infty} \left(\langle f_T, \psi_j \rangle_Y - \lambda_j \eta_j \langle \pi, \chi_j \rangle_W \right)^2 \tag{1.8}$$

[2]An alternative argument using characteristic functions instead of the singular value decomposition is provided in the subsequent Remark 3 for interested readers.

The final step is to substitute $f_T = K_1\pi_0$ and then substitute in the singular value decomposition in for K_1, keeping in mind that the ψ_j are orthonormal. The population objective function can now be expressed as the following series.

$$Q_0(\pi) = \sum_{j=0}^{\infty} \eta_j^2 \lambda_j^2 \langle \chi_j, \pi_0 - \pi \rangle_W^2 \tag{1.9}$$

The singular values η_j, λ_j are all strictly positive even though they decay. Since all of the summands are squared real numbers, $Q_0(\pi) = 0$ if and only if $\langle \chi_j, \pi_0 - \pi \rangle_W = 0$ for all j. But, since the generalized Hermite polynomials χ_j form a complete basis, any function in $\mathscr{L}_W$ that is orthogonal to all of the χ_j must be equal zero almost everywhere (Johnston, 2014). This means that any π that satisfies $Q_0(\pi) = 0$ must equal π_0 almost everywhere. If two probability density functions agree almost everywhere, then they are the same density. This gives identification of π_0, which is stated formally in Proposition 1. Its proof formalizes the argument sketched in the preceding paragraphs.

Proposition 1. *If Assumption 1 holds, then $Q_0(\pi)$ is uniquely minimized over $\mathscr{L}_Y$ by π_0.*

Proof: Section 1.13.4.

Remark 3 provides an alternative proof of identification that does not use the singular value decomposition. While the argument in the remark is more traditional and accessible, its technique does not get us much farther than identification.

Remark 3. Identification can also be shown with characteristic functions instead of the singular value decomposition. Briefly, $Q_0(\pi)$ will be zero if and only if the following unweighted integral is zero.

$$\int_{-\infty}^{\infty} \left(f_T(t) - \left(\int_{-\infty}^{\infty} \varphi(t-h)\pi(h)dh \right) \right)^2 dt = 0 \iff Q_0(\pi) = 0$$

By the Plancherel Theorem, the integrand can be replaced with its Fourier transform, i.e. the difference in characteristic functions of the two distributions. By the Convolution

Theorem, the characteristic function of a convolution of two densities is the pointwise product of the characteristic functions. This yields the following integral in terms of the characteristic functions $\zeta_{\pi_0}, \zeta_{\pi}$

$$\int_{-\infty}^{\infty} e^{-\omega^2} \left(\zeta_{\pi_0}(\omega) - \zeta_{\pi}(\omega)\right)^2 d\omega = 0 \iff Q_0(\pi) = 0$$

In order for the integral on the left-hand side to be zero, the characteristic functions must be equal almost everywhere, which means that π_0, π are the same distribution.

1.4.2 Estimation

The meta analyst observes a sample of n t-scores $t_{i,k}$ indexed by $i \in \{1, \cdots n\}$ reported by m studies indexed by $k \in \{1, \cdots m\}$. Let the number of t-scores per study be uniformly bounded. t-scores are independent across studies, but not necessarily independent within a study. The k subscript will sometimes be suppressed. For now there is no selective reporting. The meta-analyst wishes to construct an estimate $\hat{\pi}_n$ by computing and minimizing a sample analogue $\hat{Q}_n(\pi)$ of the population objective $Q_0(\pi)$ as in Newey and McFadden (1994). They then wish to use $\hat{\pi}_n$ to compute Δ_c.

Even though π_0 is identified, estimating it is a severely ill-posed inverse problem because the singular values $\lambda_j \eta_j$ decay to zero exponentially fast. This means that information about the high-frequency components of π_0 is severely attenuated in the distribution of T. The rate at which is is possible to estimate π_0 depends on how we measure the difference between the estimated density and the true density. If we take this difference to be the $\mathscr{L}_\infty$ norm, then the minimax rate is known to be logorithmic in n (Fan, 1991). This is extremely slow.

Fortunately the $\mathscr{L}_\infty$ norm is too high a standard for the meta-analyst's purposes because they only need to estimate the features of π_0 that matter for counterfactual power. I now propose a new metric that weights up the features of π_0 that matter for the

estimand and weights down features that do not matter. Let $\rho_c(\cdot,\cdot) : \mathscr{L}_W \times \mathscr{L}_W \to \mathbb{R}$ be defined as the following:

$$\rho_c(\pi_1,\pi_2) \equiv \sum_{j=0}^{\infty} \eta_j \left| \langle \chi_j, \pi_1 \rangle_W - \langle \chi_j, \pi_2 \rangle_W \right| \tag{1.10}$$

This is connected to Δ_c. Lemma 4 below says that if some sequence π_n is converging to π_0 at a certain rate in $\rho_c(\cdot,\cdot)$, then the meta-analyst can compute an estimate of Δ_c that converges at the same rate by plugging π_n into Lemma 3.

Lemma 4. *For any random sequence $\pi_n \subseteq \mathscr{L}_W$, if $\rho_c(\pi_0,\pi_n) = O_p(r_n)$, then*

$$\Delta_c - \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_n \rangle_W \left(\lambda_j \int_{-cv}^{cv} \phi_j(t)dt - \int_{-cv/c}^{cv/c} \phi_j(t)dt \right) = O_p(r_n)$$

Proof: Section 1.13.5.

The next step is to propose an estimator $\hat{\pi}_n$ for π_0. My estimator will be the minimizer of a sample analogue to the population objective function. There are many ways to construct the sample version of Q_0. The following method draws a direct connection to the singular value decomposition, and is easy to regularize and compute. Starting from Equation 1.8, notice that the inner products $\langle f_T, \psi_j \rangle_Y$ are integrals over the probability density f_T. Integrals over probability densities are expectations. So the population objective can be expressed in terms of population moments:

$$Q_0(\pi) = \sum_{j=0}^{\infty} \left(E\left[\psi_j(T) \varphi_{\sigma_Y^2}(T) \right] - \lambda_j \eta_j \langle \pi, \chi_j \rangle_W \right)^2 \tag{1.11}$$

This immediately suggests a sample analogue to the population objective where sample averages are plugged in place of the expectations. But some care is needed. Instead of summing over all j I will regularize the sample objective by summing only up to the integer J_n which grows with n. This smoothing method is called "spectral cutoff."

The sample objective $\hat{Q}_n(\pi)$ can now be expressed as:

$$\hat{Q}_n(\pi) = \sum_{j=0}^{J_n} \left(\frac{1}{n} \sum_{i=1}^{n} \psi_j(t_i) \varphi_{\sigma_Y^2}(t_i) - \lambda_j \eta_j \langle \pi, \chi_j \rangle_W \right)^2 \tag{1.12}$$

The fact that the χ_j form an orthonormal basis for $\mathscr{L}_W$ guarantees that $\hat{Q}_n$ can be minimized at zero. Strictly speaking, spectral cutoff guarantees that there are infinitely many sample minimizers, but this does not matter. I define my estimator $\hat{\pi}_n$ as the minimizer of smallest norm. This is easily computed because it is a linear combination of known Hermite polynomials where the weights are sample averages.

$$\hat{\pi}_n(h) = \sum_{j=0}^{J_n} \frac{\frac{1}{n} \sum_{i=1}^{n} \psi_j(t_i) \varphi_{\sigma_Y^2}(t_i)}{\eta_j \lambda_j} \chi_j(h) \tag{1.13}$$

The estimation error of $\hat{\pi}_n$ in the metric $\rho_c(\cdot,\cdot)$ is straightforward to compute. It is composed of two sums. The first sum is random sampling error which comes from the difference between the sample means and their expectations. The second sum is the deterministic "regularization bias" that is the price paid for cutting the sum off at J_n.

$$\rho_c(\pi_0, \hat{\pi}_n) = \underbrace{\sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_j(t_i) \varphi_{\sigma_Y^2}(t_i) - E\left[\psi_j(T) \varphi_{\sigma_Y^2}(T) \right] \right|}_{\text{Sampling Error}} + \underbrace{\sum_{j=J_n+1}^{\infty} \eta_j \left| \langle \pi_0, \chi_j \rangle_W \right|}_{\text{Reg. Bias}} \tag{1.14}$$

To see why cutting the sum off at J_n is necessary, consider the variance of the terms inside the first sum. Since λ_j is decaying to zero, if J_n grows too quickly (or is infinite) then the variance of $\hat{\pi}_n$ can explode. The expectation of the square of the sampling error can be bounded by studying the rate of decay of the λ_j and by uniformly bounding the functions $\psi_j(t)\varphi_{\sigma_Y^2}(t)$ over all j,t (Indritz, 1961). This argument yields a

bound on the sampling error.

$$\underbrace{\sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n} \sum_{i=1}^{n} \psi_j(t_i) \varphi_{\sigma_Y^2}(t_i) - E\left[\psi_j(T) \varphi_{\sigma_Y^2}(T) \right] \right|}_{\text{Sampling Error}} = O_p\left(n^{-\frac{1}{2}} \lambda_{J_n}^{-1} \right)$$

The cost of cutting the sum off at J_n is a (vanishing) regularization bias term. This bias term can be bounded using the fact that since π_0 is a probability density function with finite height, all of its inner products can be uniformly bounded. This means that the sum of regularization bias terms is bounded by a geometric series which is the same order of its first summand.

$$\underbrace{\sum_{j=J_n+1}^{\infty} \eta_j \left| \langle \pi_0, \chi_j \rangle_W \right|}_{\text{Reg. Bias}} = O(\eta_{J_n})$$

The meta-analyst wishes to tighten the regularization parameter at a rate that maximizes the speed at which $\rho_c(\pi_0, \hat{\pi}_n)$ converges in probability to zero. This means balancing bias and variance. Surprisingly, the rate-optimal choice of smoothing parameter does not depend on c. This is a consequence of the spectral cutoff regularization method. Theorem 1 gives the optimized rate. At this optimized rate the regularization bias is possibly of the same order as the sampling error.

Theorem 1. *Let Assumption 1 hold and assume there is no publication bias. Then, for any constant $\sigma_Y^2 > 0$ and any sequence $J_n \subseteq \mathbb{N}$ chosen by the meta-analyst,*

$$\rho_c(\pi_0, \hat{\pi}_n) = O_p\left(n^{-\frac{1}{2}} \lambda_{J_n}^{-1} + \eta_{J_n} \right)$$

If the meta-analyst chooses J_n, σ_Y^2 such that $n^{1/2} \left(\frac{\sigma_Y^2}{\sigma_Y^2+1} \right)^{J_n/2}$ converges to a positive number,

then:

$$\rho_c(\pi_0, \hat{\pi}_n) = O_p\left(n^{-\frac{q}{2}}\right), \qquad \textit{where } q \equiv \log\left(\frac{1+\sigma_Y^2}{1+\sigma_Y^2-c^{-2}}\right) / \log\left(\frac{1+\sigma_Y^2}{\sigma_Y^2}\right)$$

Proof: Section 1.13.6.

The rate of convergence may seem complicated but it contains several insights. If the meta-analyst wishes to estimate Δ_c without any sample size change, then they set $c = 1$. Then $q = 1$ and the estimator achieves the parameteric rate $n^{-1/2}$ in ρ_c. As the meta-analyst increases c, the rate of convergence slows down. This illustrates that counterfactual power resulting from a larger sample size increase is harder to estimate. The intuition is the following. The larger c is, the greater the difference between $h = 0$ and h close to zero. Therefore for large c, the high-frequency components of π_0 matter more. Since the high-frequency components are harder to estimate, the rate of convergence slows down.

The meta-analyst now wishes to estimate Δ_c by plugging $\hat{\pi}_n$ directly into Lemma 3. I call this estimator $\hat{\Delta}_{c,n}$. It can be expressed compactly as a weighted sum of sample means which makes $\hat{\Delta}_{c,n}$ fast to compute.

$$\hat{\Delta}_{c,n} = \sum_{j=0}^{J_n} \left(\frac{1}{n}\sum_{i=1}^{n} \psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)\right)\left(\int_{-cv}^{cv} \phi_j(t)dt - \frac{1}{\lambda_j}\int_{-cv/c}^{cv/c} \phi_j(t)dt\right)$$

Combining Theorem 1 with Lemma 4, it is immediate that $\hat{\Delta}_{c,n}$ converges to Δ_c at the same rate as $\hat{\pi}_n$ converges to π_0 in the metric ρ_c.

Corollary 1. *Under the same conditions as Theorem 1,*

$$\Delta_c - \hat{\Delta}_{c,n} = O_p\left(n^{-\frac{1}{2}}\lambda_{J_n}^{-1} + \eta_{J_n}\right)$$

Remark 4. One virtue of $\hat{\Delta}_{c,n}$ is that the signs of the observations t_j do not matter. This

is useful because in many meta-data sets the t-scores have all been reported as positive. To see why the estimator is unaffected by signs, look at the polynomial definitions in Appendix 1.12.1 and notice that if j is odd then the polynomial $\phi_j(t), \psi_j(t)$ are odd and if j is even then $\phi_j(t), \psi_j(t)$ are even. Since the integrals in $\hat{\Delta}_{c,n}$ over $\phi_j(t)$ are symmetrical about zero, then the summands where j is odd are zero. For the rest of the summands, the functions $\psi_j(t)$ are even so the signs of the data points t_i do not matter.

Remark 5. The researcher may wish to optimize the rate of convergence by choosing the tuning parameter σ_Y^2 that makes q as large as possible. By L'Hopital's Rule, as σ_Y^2 goes to infinity, q converges to c^{-2}, making the rate approach $n^{-\frac{1}{2c^2}}$. But, taking σ_Y^2 to infinity also causes the optimal J_n to grow to infinity for a fixed n, which imposes computational burdens. The choice of σ_Y^2 is therefore a balance between rate of convergence and computability. In my simulations and empirical application I always choose $\sigma_Y^2 = 1$ for reproducibility.

Remark 6. There are other methods of regularizing deconvolutions. For example Carrasco and Florens (2011) use the Tikhonov or $||\cdot||_2$ regularization which is the same technique used in ridge regression. My choice of the "spectral cutoff" regularization method is useful for this particular problem because when I add publication bias later on, it will allow me to concentrate the sample objective function. This would not be so straightforward for Tikhonov or other methods. Moreover, in this particular problem the smoothing parameter for spectral cutoff does not depend on c. This reduces the meta-analyst's researcher degrees of freedom.

1.5 Publication Bias

It is not realistic in practice to assume that every t-score computed by an experimenter is reported with equal probability. After t-scores are computed, only a subset may be published. There is growing evidence that t-scores in economics and the social

sciences are selected for publication partially on the basis of whether they cross certain significance thresholds (Brodeur et al., 2016, 2020; Andrews and Kasey, 2019; Franco et al., 2014). In this section I add publication bias to the problem. I show that π_0 is still identified and consistently estimable by adding a term to the objective function that penalizes discontinuities in the t-curve.

1.5.1 Identification under Publication Bias

This paper approaches publication bias by specifying a parametric model of selective reporting. Let R be the event indicating that the t-score T was reported. Let the conditional probability of reporting be equal to:

$$Pr(R \mid T = t) = w_{\theta_0}(t) \tag{1.15}$$

where the form of the function $w_\theta(t)$ is known to the meta-analyst and the parameter θ_0 is an unknown member of the compact set $\theta_0 \in C \subseteq \mathbb{R}^K$. I make several assumptions about the publication bias model. Assumption 2 below requires that $w_\theta(t)$ be bounded away from zero and from infinity. Without an assumption like this, π_0 is not necessarily identified.

Assumption 2. *There is a constant $\overline{\theta} > 1$ such that $\frac{1}{\overline{\theta}} \leq w_\theta(t) \leq \overline{\theta}$ for all $t \in \mathbb{R}$ and all $\theta \in C$.*

Under publication bias the meta-analyst observes only T conditional on $R = 1$. This is problematic because the expectations $E[\psi_j(T)\varphi_{\sigma_Y}(T)]$ from Equation 1.8 are no longer directly available in population. Instead the population distribution available to the meta-analyst is the conditional distribution $T \mid R$. If the meta-analyst knew θ_0, then to recover an expectation over T, they could weight the $T|R$ and take the weighted expectation. Lemma 5 computes this weighting.

Lemma 5. *If Assumptions 1 and 2 hold, then:*

$$E[\psi_j(T)\varphi_{\sigma_Y}(T)] = E\left[\frac{\psi_j(T)\,\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)}\,|\,R\right]/E\left[\frac{1}{w_{\theta_0}(T)}\,|\,R\right]$$

Proof: Section 1.13.7.

The meta-analyst must estimate θ_0. For θ_0 to be identified, publication bias must always distort the distribution of t-scores in a way that could never happen "naturally." I now add an assumption on the shape of $w_\theta(t)$ to this effect. In the absence of publication bias, T must always have a continuous probability density function. Most models of publication bias specify that reporting decisions are made on the basis of statistical significance, i.e. on whether T crosses certain thresholds. If publication bias introduces discontinuities into the density f_T, then it is possible to identify it. Assumption 3 stipulates that publication bias "reveals itself" through discontinuities at a finite set of known points. These discontinuities are sometimes called "Caliper Gaps" and they have been well studied by others (Gerber and Malhotra, 2008; Elliott et al., 2022; Kudrin, 2023).

Assumption 3. *There is a finite set $\mathscr{X} \subseteq \mathbb{R}$ and a constant $L > 0$ such that for all $\theta_1, \theta_2 \in C$ the function $\frac{w_{\theta_1}(t)}{w_{\theta_2}(t)}$ is uniformly continuous in t on $\mathscr{X}^c$ and*

$$\sup_{t\in\mathbb{R}} \frac{1}{L}\left|\frac{1}{w_{\theta_1}(t)} - \frac{1}{w_{\theta_2}(t)}\right| \leq ||\theta_1 - \theta_2||_\infty \leq \max_{x\in\mathscr{X}} L \lim_{\varepsilon\to 0}\left|\frac{w_{\theta_1}(x+\varepsilon)}{w_{\theta_2}(x+\varepsilon)} - \frac{w_{\theta_1}(x-\varepsilon)}{w_{\theta_2}(x-\varepsilon)}\right|$$

Assumption 3 is useful for the following reason. Suppose that the meta analyst has a guess θ for θ_0 and attempts to remove the publication bias by reweighting the density $f_{T|R}$ by $1/w_\theta$. Then the meta-analyst can tell how close they got to f_T by checking the magnitude of the largest discontinuity left in the reweighted density. If the meta-analyst has successfully removed every discontinuity, then $1/w_\theta = 1/w_{\theta_0}$ and they have recovered f_T. Example 2 shows how the "illustrative example" of publication bias from Andrews

and Kasey (2019) satisfies Assumption 3. This will be the model of publication bias that I use in all of the simulations and empirical applications later on.

Example 2. *Here is a simple parametric model of publication bias that satisfies all of the assumptions in this paper. Suppose that the probability of publication depends only on whether* $|T|$ *falls above the critical value* 1.96*. The conditional probability ratio is equal to the scalar* $\theta_0 = \frac{P(R\,||T|<1.96)}{P(R\,||T|\geq 1.96)} \in \left(\frac{1}{\overline{\theta}}, \overline{\theta}\right)$*. This means that the weighting function is:*

$$w_{\theta_0}(t) = \theta_0 \mathbf{1}\{|t| < 1.96\} + \mathbf{1}\{|t| \geq 1.96\}$$

In this case the set $\mathscr{X}$ *contains just* $\{1.96, -1.96\}$*, the set C is* $\left(\frac{1}{\overline{\theta}}, \overline{\theta}\right)$ *and the Lipschitz constant L is* $\overline{\theta}$*. This model is used by Andrews and Kasey (2019).*

These assumptions are enough to identify π_0, and θ_0. To show this, consider the following publication-bias-aware population objective function $Q_0(\pi, \theta)$. This is similar to the objective in the absence of publication bias with two differences. First, the weighted expectations conditional on R have been substituted for the unweighted unconditional expectations. Second there is now a second sum that penalizes Caliper discontinuities in the weighted density.

$$\begin{aligned} Q_0(\pi, \theta) \equiv \sum_{j=0}^{\infty} \left(E\left[\frac{\psi_j(T)\,\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)} \,|\, R\right] / E\left[\frac{1}{w_{\theta_0}(T)} \,|\, R\right] - \lambda_j \eta_j \langle \pi, \chi_j \rangle_W \right)^2 \\ + \sum_{x \in \mathscr{X}} \lim_{\varepsilon \to 0} \left(\frac{f_{T\,|R}(x+\varepsilon)}{w_\theta(x+\varepsilon)} - \frac{f_{T\,|R}(x-\varepsilon)}{w_\theta(x-\varepsilon)} \right)^2 \end{aligned}$$

Proposition 2 says that π_0, θ_0 are identified when the publication bias is correctly specified. To see why, notice that Assumption 3 guarantees that the second sum is zero if and only if $\theta = \theta_0$. Concentrating at the true θ_0 yields $Q_0(\pi, \theta_0)$ which is uniquely minimized over π by π_0 by Proposition 1. So π_0 and θ_0 are identified.

Proposition 2. *If Assumptions 1, 2, and 3 hold and publication bias follows Equation 1.15, then $Q_0(\pi, \theta)$ is uniquely minimized over $\mathscr{L}_Y$ by π_0, θ_0.*

Remark 7. Andrews and Kasey (2019) identify publication bias by assuming that each study's standard error and estimand are independent. This is defensible in the context of a specific and narrow collection of studies that are all estimating similar estimands with similar methods. But for the applications in this paper this assumption is not defensible because I examine literatures that contain highly heterogeneous studies. For such literatures, correlation between standard errors and estimands can occur in many ways, e.g. when different studies report results using different units.

Remark 8. The term penalizing discontinuities in derivatives can actually be dropped from $Q_0(\pi, \theta)$ and identification can still hold. However, the regularized plug-in estimator suggested by this alternative population objective seems to perform poorly in simulation so I do not study it in this paper.

1.5.2 Estimation under Publication Bias

The meta-analyst estimates π_0, θ_0 jointly by minimizing a sample analogue of the population objective function. As in the case without publication bias, the meta-analyst obtains the sample objective by plugging sample averages in place of expectations and regularizes using the spectral cutoff method by summing only over J_n terms.

The sample analogue to the penalty for Caliper discontinuities is constructed in the following way. The meta-analyst first chooses a bin width $\varepsilon_n > 0$ and computes a histogram from the sample of published t-scores. The height of the histogram bars is divided by ε_n so that the histogram approximates the density $f_{T|R}$. Then the heights of the histogram bars are weighted by $w_\theta(t)$ such that weighted histogram would approach continuous function as $\varepsilon_n \to 0$ if θ were equal to θ_0. Finally, the differences between pairs of histogram bars straddling the members of $\mathscr{X}$ are penalized in the objective. This

yields the following sample objective:

$$\hat{Q}_n(\pi,\theta) \equiv \sum_{j=0}^{J_n} \left(\frac{\frac{1}{n}\sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_\theta(t_i)}}{\frac{1}{n}\sum_{i=1}^n \frac{1}{w_\theta(t_i)}} - \eta_j \lambda_j \langle \chi_j, \pi \rangle_W \right)^2 + \sum_{x \in \mathscr{X}} \left(\frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i \in (x, x+\varepsilon_n]\}}{w_\theta(x+\varepsilon_n)} - \frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i \in (x-\varepsilon_n, x]\}}{w_\theta(x-\varepsilon_n)} \right)^2$$

There is now a third tuning parameter ε_n to choose. If ε_n is large, then the penalty is likely to be large even when we input the correct parameter $\theta = \theta_0$. This may lead to substantial bias. If ε_n is small, then the penalty term will have high variance and be unreliable. The optimal rate at which to decrease ε_n as n increases turns out to be $n^{-1/3}$ which is the same as the optimal rate for pointwise convergence of the histogram. In Section 2.7, I provide more specific recommendations.

The key to estimation under publication bias is that π can easily be concentrated out of the sample objective function. To avoid confusion with the previous estimator I call the estimator under publication bias $\hat{\pi}_n^{pb}$. Once again because there are infinitely many π that set the sample objective to zero, I define $\hat{\pi}_n^{pb}$ as the function of smallest norm which is expressed below in terms of $\hat{\theta}_n$.

$$\hat{\pi}_n^{pb} = \frac{1}{\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{j=0}^{J_n} \frac{\frac{1}{n}\sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)}}{\eta_j \lambda_j} \chi_j$$

Since π has been concentrated out of the sample objective and everything but the Caliper penalty sum has been set to zero, minimization proceeds by minimizing the penalty over θ. The sample minimizer $\hat{\theta}_n$ expressed below can be plugged back into the previous equation to obtain $\hat{\pi}_n^{pb}$.

$$\hat{\theta}_n = \min_\theta \sum_{x \in \mathscr{X}} \left(\frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i \in (x, x+\varepsilon_n]\}}{w_\theta(x+\varepsilon_n)} - \frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i \in (x-\varepsilon_n, x]\}}{w_\theta(x-\varepsilon_n)} \right)^2$$

Theorem 2 shows the consistency of $\hat{\pi}_n^{pb}$. The rate of convergence has slowed down to $n^{-q/3}$ compared to the case without publication bias. This is because a small change in the histogram near a point in x can change the weight that every t_i gets in each of the sample averages in the sample objective. The previous remarks 4, 5, and 6 also apply here.

Theorem 2. *Let Assumptions 1, 2, and 3 hold and assume that publication bias follows Equation 1.15. If the meta-analyst chooses* $\varepsilon_n \propto n^{-\frac{1}{3}}$

$$\rho_c\left(\pi_0, \hat{\pi}_n^{pb}\right) = O_p\left(\lambda_{J_n}^{-1} n^{-\frac{1}{3}} + \eta_{J_n}\right)$$

If the meta-analyst also chooses J_n, σ_Y^2 *such that* $n^{1/3}\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)^{J_n/2}$ *converges to a positive number, then:*

$$\rho_c\left(\pi_0, \hat{\pi}_n^{pb}\right) = O_p\left(n^{-\frac{q}{3}}\right), \qquad \text{where } q \equiv \log\left(\frac{1+\sigma_Y^2}{1+\sigma_Y^2-c^{-2}}\right) / \log\left(\frac{1+\sigma_Y^2}{\sigma_Y^2}\right)$$

Proof: Section 1.13.9.

Just as in the previous section, the meta-analyst can proceed to estimate the unconditional counterfactual power by plugging $\hat{\pi}_n$ into Lemma 3. Since the sample objective function can be concentrated, it is possible to express the estimator $\hat{\Delta}_{c,n}^{pb}$ in terms of sample averages just as in the case without publication bias.

$$\hat{\Delta}_{c,n}^{pb} \equiv \frac{\sum_{j=0}^{J_n}\left(\int_{-cv}^{cv}\phi_j(t)dt - \frac{1}{\lambda_j}\int_{-cv/c}^{cv/c}\phi_j(t)dt\right)\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}}$$

Combining Theorem 2 and Lemma 4, gives us the rate of consitency for Δ_c.

Corollary 2. *Under the same conditions as Theorem 2,*

$$\Delta_c - \hat{\Delta}_{c,n}^{pb} = O_p\left(n^{-\frac{q}{3}}\right)$$

1.6 Inference

Next I show asymptotic normality of $\hat{\Delta}^{pb}_{n,c}$ and derive a consistent variance estimator. Asymptotic normality is shown using several Taylor approximations. I will need to assume that the estimator does not converge so fast that the Taylor residuals become leading terms.Since the regularization bias can be of the same order as the sampling error, I will also need to center $\hat{\Delta}^{pb}_{n,c}$ on a deterministic sequence $\tilde{\Delta}^{pb}_{n,c}$ that converges to Δ_c but is not necessarily equal to it.[3]

1.6.1 Asymptotic Normality

Publication bias makes the estimator nonlinear. This is dealt with in the usual way using Taylor approximations. In order for Taylor approximations to be valid, $w_\theta(t)$ needs to be sufficiently smooth in θ. This motivates Assumption 4. It says that the gradient of $w_\theta(t)$ in θ is Lipschitz continuous in θ.

Assumption 4. *There is an $L > 0$ such that for all $\theta_1, \theta_2 \in C$, the $K \times 1$ gradients of $w_\theta(t)$ satisfy:*

$$\left|\left|\nabla_\theta \frac{1}{w_{\theta_1}(t)} - \nabla_\theta \frac{1}{w_{\theta_2}(t)}\right|\right|_\infty \leq L||\theta_1 - \theta_2||_\infty$$

Regularization bias will affect the centering of the estimator. The centering term is non-negligible because $\hat{\Delta}^{pb}_{n,c}$ contains smoothing bias from two different sources. First, since only the first J_n terms of the singular value decomposition are used, the set the rest of the terms are set to zero and this incurs regularization bias. Second, since the meta-analyst only observes a sample of t-scores, they estimate the discontinuities in $f_{T|R}$ by looking in a window of width ε_n on either side of each possible point of discontinuity. Since the PDF can change over this interval, smoothing bias is incurred here as well.

I call the centering term $\tilde{\Delta}^{pb}_{c,n}$. Its expression is lengthy and is relegated to Appendix

[3]Ongoing work aims to lower bound the variance of the estimator, allow for undersmoothing, and remove these caveats.

1.12.2. The important point for now is that $\tilde{\Delta}^{pb}_{c,n}$ converges to Δ_c at least as fast as $\hat{\Delta}^{pb}_{c,n}$ converges. Lemma 6 proves this.

Lemma 6. *If Assumptions 1-4 hold, then:*

$$\Delta_c - \tilde{\Delta}^{pb}_{c,n} = O\left(\eta_{J_n} + \varepsilon_n\right)$$

If the meta-analyst chooses $\varepsilon_n = O\left(n^{-1/3}\right)$ *and* $n^{1/3}\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)^{J_n/2} = O(1)$*, then*

$$\Delta_c - \tilde{\Delta}^{pb}_{c,n} = O\left(n^{-q/3}\right)$$

Proof: Section 1.14.2.

After centering properly, it is possible to linearize the estimator $\hat{\Delta}^{pb}_{c,n}$. Our goal is to construct a triangular array of sample means. Lemma 7 below uses Taylor's Theorem several times to rewrite the estimator as a sample mean plus a vanishing term. The function $Z_{n,J_n}(t)$ is deterministic. Since it is lengthy to write down, its expression is relegated to Appendix 1.12.3.

Lemma 7. *If Assumptions 1-4 hold, and the meta-analyst chooses* $\varepsilon_n \propto n^{-1/3}$*, then there exist deterministic functions* $Z_{n,J_n} : \mathbb{R} \to \mathbb{R}$ *such that (i)* $\max_n \sup_{t\in\mathbb{R}} n^{-1/3}\lambda_{J_n}|Z_{n,J_n}(t)| < \infty$ *and (ii)* $\max_n n^{-1/3}\lambda_{J_n}^2 Var\left(Z_{n,J_n}(T) \mid R\right) < \infty$ *and*

$$\hat{\Delta}^{pb}_{c,n} - \tilde{\Delta}_{c,n} = \frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i) - E\left[Z_{n,J_n}(T)\right] + O_p\left(n^{-2/3}\lambda_{J_n}\right)$$

Proof: Section 1.14.3.

Next I show that the triangular array of sample means $\frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i)$ is asymptotically normal. In order for this to guarantee that the estimator $\hat{\Delta}^{pb}_{c,n}$ is itself also normal, the sample means must dominate the Taylor residuals. To guarantee this I add the next

assumption. Assumption 5 says that the estimator does not converge much faster than its guaranteed rate.

Assumption 5.

$$\liminf_{n\to\infty} n^{5/6}\lambda_{J_n}^2 Var\left(\hat{\Delta}_{c,n}\right) = \infty$$

To show asymptotic normality of the triangular array of sample sums $\frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i)$ I use a Lindeberg-Feller type Central Limit Theorem. The key step is to verify the Lyapunov Condition. While verifying this condition is a common tactic in ill-posed problems, my argument is quite different than Carrasco and Florens (2011). Using the bounds on the magnitude and variance of the Z_{n,J_n} from Lemma 7, we can verify that Assumption 5 is sufficient to guarantee the following Lyapunov Condition that uses the fourth moments.

$$\frac{E\left[\left(Z_{n,J_n}(T)\right)^4\right]}{nE\left[\left(Z_{n,J_n}(T)\right)^2\right]^2} \to 0$$

Since each observation of $T|R$ is identically distributed and independent of all but at most C other observations, the dependence is weak and the sample means are asymptotically normal. Theorem 12 shows that the assumptions so far are enough to satisfy all of the hypotheses of the central limit theorem in Theorem 2.1 of Neumann (2013).

Theorem 3. *If Assumptions 1-5 hold and the meta-analyst chooses $\varepsilon_n \propto n^{-1/3}$, then:*

$$\frac{\hat{\Delta}_{c,n} - \tilde{\Delta}_{c,n}}{\sqrt{Var(\hat{\Delta}_{c,n})}} \to_d N(0,1)$$

Proof: Section 1.14.1.

1.6.2 Variance Estimation

The next step is to derive an estimator consistent for the variance of $\frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i)$. Under Assumption 5, this sample mean dominates the other random terms in $\hat{\Delta}_{n,c}^{pb}$ and it

is enough to estimate its variance. In other words, Assumption 5 guarantees that the ratio of the variances converges to one:

$$\frac{Var(\hat{\Delta}_{n,c})}{Var\left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i)\right)} \to 1$$

The variance of the sample mean in the denominator is straightforward to derive. The variance of the sum is the sum of the covariances. Since two t-scores drawn from different studies are independent, the covariances are zero across studies. Let the symmetric $n \times n$ adjacency matrix Λ equal 1 when two t-scores are from the same study. This gives us the following expression for the variance:

$$Var\left(\frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} \Lambda_{ik}\text{Cov}\left(Z_{n,J_n}(t_i), Z_{nJ_n}(t_k)\right)$$

Since the functions Z_{n,J_n} depend on the true θ_0 and π_0, the meta-analyst does not know them. But there is a sample version $\hat{Z}_{n,J_n}$ that the meta-analyst does observe and can be used for variance estimation. Its expression is once again very lengthy and is relegated to Appendix 1.12.3. Theorem 4 guarantees that variance estimation is consistent by proving that $\hat{Z}_{n,J_n}(t)$ converge uniformly to $Z_{n,J_n}(t)$ fast enough for a variance estimator to converge faster than the variance itself decays to zero.

Theorem 4. *If Assumptions 1-5 hold and the meta-analyst chooses $\varepsilon_n \propto n^{-1/3}$ then:*

$$\frac{\hat{\Delta}_{c,n} - \tilde{\Delta}_{c,n}}{\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n} \Lambda_{ik}\hat{Z}_{n,J_n}(t_i)\hat{Z}_{nJ_n}(t_k)}} \to_d N(0,1)$$

Proof: Section 1.14.4.

Theorem 4 says that the meta-analyst can construct valid confidence intervals covering the centering sequence $\tilde{\Delta}_{c,n}$. Lemma 3 showed that the centering sequence converges to Δ_c at the same rate as the variance decays. In theory, this could affect the

coverage of the confidence intervals for Δ_c itself. A natural solution is to "undersmooth" or to change J_n, ε_n more quickly than the optimal rate in order to force $\tilde{\Delta}_{c,n}$ to converge to Δ_c faster than the variance decays. Ongoing work aims to lower bound the variance of the estimator to allow for undersmoothing. The simulations and empirical applications in this paper choose not to undersmooth.

1.7 Simulations

To compute $\hat{\Delta}^{pb}_{n,c}$, the meta-analyst must choose three tuning parameters: ε_n, J_n, and σ_Y. To guarantee the optimized rates of convergence in Theorems 1 and 2 the meta-analyst sets $\varepsilon_n = Cn^{-1/3}$ and $J_n = \log\left(Dn^{-1/3}\right)/\log\left(\sigma_Y^2/\left(1+\sigma_Y^2\right)\right)$ where $C, D > 0$. This means that the meta-analyst's choice is actually over the triple of constants $\{C, D, \sigma_Y\}$. The theorems in the preceding sections provide no specific guidance on how these constants are to be chosen and we must turn to simulations.

I recommend that the meta-analyst should always at least disclose results using the following tuning parameters: $C = 2$, $D = 10^{-4}$, and $\sigma_Y = 1$. These choices of tuning parameters yield good confidence interval coverage of Δ_c in simulation for a very wide variety of data generating processes. I have found that using a larger D sometimes leads to reduced coverage and that decreasing D widens confidence sets with no gain. Choosing a larger σ_Y increases computational costs with no obvious gain in simulation.

I generate simulated data in the following way. First I use the simple model of publication bias from Example 2 where t-scores are reported with probability θ_0 if they do not clear 1.96 and are reported with certainty otherwise. This matches the illustrative example from Andrews and Kasey (2019). In the simulations in this section I set $\theta_0 = 0.9$ but the results are not sensitive to this.

The small sample performance of the confidence intervals could depend on the distribution of true effects π_0. I report simulations for data generating processes where

I expect coverage to be poor. Theory predicts that coverage could be low in three situations. First, if the distribution π_0 is not very smooth, then the coefficients $\langle \chi_j, \pi_0 \rangle$ decay slowly in j and regularization bias is large. Similarly, if π_0 has large maximum height, then $||\pi_0||_W$ is large which could also increase regularization bias. Third, if the density f_T is close to zero or has steep slope at the critical threshold 1.96 then $\hat{\theta}_n$ is poor and this could in theory affect coverage of Δ_c. I report simulations against five different distributions π_0, four of which meet some or all of these criteria.

- First consider a situation where every true effect is very close to zero. I call this the "Null" density and specify it to be $h \sim \text{Unif}\left(-10^{-3}, 10^{-3}\right)$. This density is tall and non-smooth so regularization bias may be high. In addition, the density is small at 1.96.

- Second consider a case where the distribution of true effects has thick tails. Outliers are common in my empirical applications and it is useful to confirm that the confidence sets are robust to extreme t-scores. I specify $h \sim$ Cauchy. This has low density at 1.96.

- Third consider a case where the distribution of h is bimodal. Brodeur et al. (2020) documents bimodality in the empirical distribution of t-scores in economics. For the bimodal distribution I specify that h is a 50% mixture between two normals. One normal is centered at zero and another at 2.8 and both have variance one. I choose this centering because when $h = 2.8$ then conditional power is close to 80% for a size 5% test which is the standard target for power calculations. This bimodal distribution models a situation where conditional power is either close to 80% or close to 5%.

- Fourth consider the distribution that brings Δ_c close to its maximum. The "Large" distribution is $h \sim \text{Unif}\left(1.96 - 10^{-3}, 1.96 + 10^{-3}\right)$ and makes the estimand $\Delta_c = 0.28$

which is large and different from the rest of the simulations. This density is tall and non-smooth so regularization bias could be high.

- The fifth distribution puts the confidence sets to a severe test. For the "Slope" distribution I set $h \sim \text{Unif}\left(0.96 - 10^{-3}, 0.96 + 10^{-3}\right)$. This increases the bias of $\hat{\theta}_n$ by maximizing the slope of the density f_T near the critical value 1.96.

I compare performance for these five choices of π_0 under two different meta-sample sizes, a small meta-sample of 50 t-scores versus a large meta-sample of 1000 t-scores.

The simulation results are reported in Table 1.1. The first column reports the meta-sample size n. The second column reports the distribution of true effects π_0 that was used to generate the data. The third and fourth columns report unconditional power and the power gain Δ_c from a counterfactual doubling of every sample size in the population of experiments. This means that $c = \sqrt{2}$. The fifth column reports the mean over 2000 simulation repetitions of the estimate $\hat{\Delta}_{c,n}$. The sixth and seventh columns compare the standard deviation of the estimator and the average standard error. The final column reports coverage of Δ_c by the 95% confidence set centered on $\hat{\Delta}_{c,n}$. Coverage is close to 95% for all of the π_0 for both large and small n. This is not necessarily the case for other choices of D and C.

1.8 Empirical Applications

I apply $\hat{\Delta}^{pb}_{c,n}$ to two real world settings. First I present an empirical exercise using data from a multi-site replication study in laboratory psychology (Klein et al., 2014). This setting is special because counterfactual power can be credibly and precisely estimated by other means. I find that the confidence sets produced by my method contain a credible alternative point estimate. This exercise provides evidence that this novel method can perform well with real data. Second I present an empirical application where I use my method to answer a broad question of practical interest: are randomized controlled trials

published in top economics journals underpowered? My method provides evidence of the opposite: the power of t-tests reported by RCTs published in these outlets are on average quite insensitive to sample size increases.

1.8.1 Replication Studies in Laboratory Psychology

While the simulations in the previous section show good coverage for the confidence intervals against a variety of artificial data generating processes, I cannot know from those results how realistic the simulations are or how well the method performs when it is used on real world data. In this section I present an exercise that takes place in a special empirical environment where it is possible to construct a credible and precisely estimated alternative point estimate for Δ_c that I can benchmark my confidence sets against.

The Many Labs systematic replication project is an ideal environment for this exercise. Klein et al. (2014) recruited 36 independent research teams who each attempted to replicate 13 effects from experimental psychology. Each team collected their own independent data and ran some or all of the thirteen experiments on their respective samples. Every sample contained at least 80 participants and many contained far more for a total of 6344. Eleven of the experiments were analyzed using t-tests of the equality in means between a treatment group and a control group and I will limit my analysis to these. This yields a meta-sample of 385 t-scores. The aim of all of the experiments run by Many Labs was to replicate effects that had been published in top psychology journals and to investigate how consistently effects could be replicated across study sites.

This environment is special because it is plausible to assume that all of the research teams were investigating the same (or very similar) true effects. A key conclusion reported by Klein et al. (2014) is that variation in the true effect size across study sites was found to be very small compared to the variation in the effect sizes across the experimental treatments. This finding is plausible because the experiments took place in

controlled laboratory environments, the researchers were all using the same protocols, and the primary aim of every experiment was to replicate existing results consistently. Given that the research teams had no incentive to selectively report their results, it is also reasonable to assume that there was no publication bias in this setting.

Assuming that the true effects did not vary across study sites makes it possible to estimate Δ_c^{COND}, which is the power gain conditional on sample realizations of h.

$$\Delta_c^{(n)} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\Pr\left(|T_c| > 1.96 \,|\, h = h_i\right) - \Pr\left(|T| > 1.96 \,|\, h = h_i\right)\right] \tag{1.16}$$

Notice that $E\left[\Delta_c^{(n)}\right] = \Delta_c$ and in large samples, $\Delta_c^{(n)} \to_p \Delta_c$. Even though $\Delta_c^{(n)}$ is not identical to Δ_c in every sample, its meaning and interpretation are similar enough to use as a benchmark.

To estimate $\Delta_c^{(n)}$ we proceed as follows. For each of the 11 experimental treatments, I take the mean of the reported effect sizes across the 36 study sites to estimate the true effect b for that treatment. Then I replace each of the 386 t-scores with the ratio $\frac{b}{s_i}$ where s_i is the standard error used to compute the ith t-score. To compute unconditional power, I plug each ratio $\frac{b}{s_i}$ into the power function for the size 5% t-test and take the mean. To compute power were the sample sizes all to have been counterfactually doubled, I plug $\sqrt{2}\frac{b}{s_i}$ into the power function instead. Taking the average yields the point estimate: $\hat{\Delta}_c^{(n)} = 0.078$. This estimation technique is a simplified and unweighted version of those used by Ioannidies et al. (2017); Arel-Bundock et al. (2022) and employs the same basic assumption. Computing the standard error under the conservative worst-case assumption that two experiments in the same lab have covariance 1 yields standard error 0.006. The alternative point estimate $\hat{\Delta}_c^{(n)}$ is quite precise.

I test the coverage of the confidence sets proposed by this paper by checking whether they contain $\hat{\Delta}_c^{(n)}$. Table 1.2 presents the results. The third row of Table 1.2 reports the 95% confidence interval in square brackets. The confidence interval contains

the benchmark value 0.078. This exercise complements the simulations by arguing that inference is likely to work reasonably well in a real-world environment.

To show that the results are not sensitive to reasonable choices of the tuning parameters, I present four specifications. In columns 1 and 2 I scale J_n, ε_n by the number of experimental study sites. In columns 3 and 4 I instead scale the tuning parameters by the total number of t-scores. In columns 1 and 3 I use the recommended values of the constants C, D from Table 1.1. In columns 2 and 4 I vary these choices to show that the confidence sets do not change much even when the changes to C, D are large.

$\Delta_c^{(n)}$ is not identical to Δ_c except in large samples. Nevertheless I argue that the benchmark is still very useful because it demonstrates that the confidence intervals produce plausible results in a well-understood empirical environment.

1.8.2 Randomized Controlled Trials in Economics

I apply $\hat{\Delta}_{c,n}^{pb}$ to answer an empirical question with policy implications: Are randomized controlled trials published in top economics journals underpowered? Experiments are lauded as the "gold standard" of empirical evidence in social science because their design allows them to credibly control the rate of type-I errors. But what about type-II errors? If the conclusions reported by influential RCTs are sensitive to reasonable increases in sample size, then funders and researchers should consider running experiments with fewer treatment arms and allocate more resources toward data collection.

This is an empirical question and the answer will depend on the population of experiments under study. Here I study the experiments that have the most influence in academic economics: those published in top journals. The data source is Brodeur et al. (2020). This meta-study examined the universe of 684 articles published in top economics journals during 2015-2018. The data contain 21,740 test statistics of which 20,419 were t-scores that I can use. All test statistics corresponded to main hypotheses of interest and excluded regression controls, robustness checks, placebo tests, and the

like. Every t-score was produced using one of four empirical methods: Randomized Controlled Trials (RCTs), Difference in Differences (DID), Discontinuity Designs (DD), and Instrumental Variables (IV).

I estimate the sensitivity of tests conducted by RCTs to hypothetical sample size increases in Table 1.3. The estimate in column 1 says that counterfactually doubling the sample sizes of every RCT published in top economics journals would only increase the expected number of t-scores clearing the critical value of 1.96 by 7.2 percentage points with standard error 2.5.

This result can be interpreted by comparing it to other empirical estimates of Δ_c. First I return to the Many Labs result from the previous section as our first benchmark. The Many Labs Δ_c should be very low because Many Labs consisted of deliberate laboratory replications of previously published results and used samples chosen to ensure adequate power. Since experiments published in top economics journals are usually novel in some way, the researcher knows less in the design phase and the sample size choice is less likely to be optimal.

Table 1.3 shows that I fail to reject the null hypothesis that Δ_c is equal across economics RCTs and the Many Labs replications $(p = 0.18)$. This is not simply a consequence of high uncertainty because the confidence intervals involved are small enough to make other meaningful comparisons. Column 2 of Table 1.3 provides an example. It reports that among tests conducted by non-RCTs, doubling sample sizes would increase power by 17.3 percentage points which is significantly different from Many Labs $(p = 0.00)$.

These results are not very sensitive to the choice of tuning parameters. Columns 3 and 4 of Table 1.3 show a robustness check where the tuning parameters J, ε are scaled by the number of articles instead of the number of t-scores. These confidence intervals are smaller but at greater risk of bias. The results are largely unchanged. The only comparison that changes is that RCTs and the Many Labs Δ_c become marginally significantly different $(p = 0.04)$. This may be because scaling by number of articles

makes confidence intervals less likely to cover. Otherwise the results are qualitatively unchanged.

Figures 1.3, 1.4, and 1.5 visualize these comparisons over many possible sample size increases c^2. Figure 1.3 shows how power climbs much faster for non-experiments than RCTs in the Brodeur et al. (2016) sample even when we increase sample sizes by less than double. Figures 1.4 and 1.5 show that the Many Labs experiments respond at a pace indistinguishable from the economics RCTs. These figures reinforce and visualize the results from the tables.

As a final benchmark, consider a literature where every experiment is run at exactly 80% power. Then we can calculate that doubling every sample size would increase power by 17.8 percentage points. We can easily reject the null hypothesis that RCTs in the Brodeur et al. (2016) data are this sensitive.

I conclude from this exercise that RCTs published in top economics journals are relatively insensitive to counterfactual sample size increases. The policy implication is that funders and researchers should generally consider devoting more resources to running more experiments or adding treatment arms rather than raising sample size standards.

There is one important caveat to mention. The estimand Δ_c is the change in unconditional power and is therefore an average over many highly heterogeneous studies. This estimand likely masks important heterogeneity. Future research could use the method in this paper to identify sub-populations of experiments that are relatively sensitive to counterfactual sample size increases.

1.9 Conclusion

This paper proposed an estimator consistent for the fraction of t-scores that would have been statistically significant had every experiment in a given collection had its

sample size counterfactually increased by a chosen factor $c > 1$. The only assumption imposed on the distribution of true effects was that it had a probability density function with bounded height. The proposed estimator is asymptotically normal and robust to simple forms of publication bias. The key insight of the paper was that even though it is not feasible to estimate the distribution of true effects π_0 it is feasible to estimate enough components of π_0 to recover counterfactual power. The key technical step was to use the spectral cutoff regularization method to concentrate the publication bias out of the sample objective function.

I provided tuning parameter recommendations that yielded good small sample coverage for the confidence sets across many simulations. An empirical exercise using the Many Labs dataset confirmed that the confidence interval contained a point estimate computed using an alternative method that the Many Labs environment made possible. In a second empirical application I found that the power of randomized trials in economics would only increase by 7.2 percentage points on average if every sample size had been doubled. I argued that this number is small. The policy implication is that funding organizations should in general fund more randomized trials rather than fewer, larger ones.

1.10 Tables for Chapter 1

Table 1.1. Simulations

Parameters				Results			
n	π_0	Un. Pwr.	Δ_c	Mean $\hat{\Delta}_{c,n}$	SD $\hat{\Delta}_{c,n}$	Mean SE $\hat{\Delta}_{c,n}$	Coverage
50	Null	0.05	0.00	0.01	0.18	0.18	0.95
50	Cauchy	0.45	0.09	0.07	0.15	0.15	0.94
50	Bimodal	0.57	0.12	0.09	0.16	0.15	0.94
50	Large	0.78	0.28	0.26	0.15	0.15	0.95
50	Slope	0.27	0.11	0.11	0.17	0.17	0.94
1000	Null	0.05	0.00	0.00	0.04	0.04	0.95
1000	Cauchy	0.45	0.09	0.08	0.03	0.03	0.96
1000	Bimodal	0.57	0.12	0.12	0.03	0.03	0.95
1000	Large	0.78	0.28	0.28	0.03	0.04	0.96
1000	Slope	0.27	0.11	0.12	0.04	0.04	0.95

Notes: Table showing simulation results. Each row is a parameterization. Each parameterization was run for 2000 repetitions. All simulations used the tuning parameters: $C = 2$, $D = 10^{-4}$, and $\sigma_Y = 1$. The first column in the table is the number of t-scores. The second is the distribution of true effects and each of these is described in Section 2.7. The third column is unconditional (mean) statistical power at the status quo. The fourth column is the increase in unconditional power resulting from doubling the sample size of every experiment, so $c = \sqrt{2}$. The Fifth and sixth columns are the mean and standard devition of the point estimates. The seventh column is the mean of the estimated standard errors and the final column is the fraction of confidence intervals that contained the true Δ_c.

Table 1.2. Empirical Exercise: Many Labs Replication Project

	By Number of t-scores		*By Number of Sites*	
	Main	Rob. Check	Main	Rob. Check
$\hat{\Delta}_c$	.005	.033	.005	.033
	(.042)	(.026)	(.039)	(.030)
	[0, .088]	[0, .083]	[0, .082]	[0, .092]
$\hat{\theta}$	.92	1.20	1.03	0.86
	(.70)	(.73)	(.30)	(.67)
D	1e-4	3e-4	1e-4	3e-4
C	2	1	2	1
J	16	15	15	13
ε	.27	.07	.61	.15
$\hat{\Delta}_c^{(n)}$	0.078	0.078	0.078	0.078
	0.006	0.006	0.006	0.006
Unconditional Power	.61	.61	.61	.61
t-scores	385	385	385	385
Sites	36	36	36	36
Treatments	11	11	11	11

Notes: Table reports estimates $\hat{\Delta}$ of gain in unconditional statistical power of a two-sided size 5% t-test resulting from counterfactually doubling every study's sample size from the status quo. Columns 1 and 2 report results where J, ε are scaled based on the number of t-scores while in columns 3 and 4 these tuning parameters are scaled based on the number of study sites. The confidence intervals in columns 1 and 2 are more conservative. Columns 1 and 3 use the preferred choice of C, D while columns 2 and 4 present a robustness check where we have significantly altered C, D. We always set $\sigma_Y = 1$. Standard errors are reported in round brackets and 95% confidence intervals are in square brackets. Standard errors clustered by study site and treatment type. The bottom five rows report the estimate and standard error of $\hat{\Delta}^{(n)}$ using the method of means, the fraction of t-scores larger than 1.96, the number of t-scores in total, the number of unique experimental research sites, and the number of unique experimental treatments. Data from Klein et al. (2014).

Table 1.3. Empirical Application: Randomized Trials in Economics

	By Number of t-scores		*By Number of Articles*	
	RCT	non-RCT	RCT	non-RCT
$\hat{\Delta}$	.072	.173	.095	.164
	(.025)	(.028)	(.020)	(.021)
	[.02, .12]	[.12, .23]	[.06, .13]	[.12, .20]
$\hat{\theta}$	1.04	.93	.73	.59
	(0.11)	(.10)	(.20)	(.17)
J	18	18	15	16
ε	.10	.08	.46	.38
$=$ RCT (p-value)		0.01		0.02
$=\Delta_c$ Many labs (p-value)	0.18	0.00	0.04	0.02
$=\Delta_c^{(n)}$ Many labs (p-value)	0.81	0.00	0.40	0.00
Status Quo Power	.38	.55	.38	.55
t-scores	7569	14171	7569	14171
Articles	80	145	80	145

Notes: Reports estimates $\hat{\Delta}$ of the gain in unconditional statistical power of a two-sided size 5% t-test resulting from counterfactually doubling every study's sample size from the status quo (c=$\sqrt{2}$). Compares t-scores testing main hypotheses from randomized trials published in top economics journals versus those reported by studies using difference in difference, regression discontinuity or instrumental variables. Tuning parameters are $D = 10^{-4}$, $C = 2$, and $\sigma_Y = 1$. Table compares results when J, ε are chosen based on the number of articles (less conservative) vs the number of t-scores (more conservative) in the meta-sample. Standard errors are reported in round brackets and 95% confidence intervals are in square brackets. Standard errors clustered by article. The bottom four rows report the outcome of a test for equality of the RCT vs non-RCT values of Δ, the fraction of t-scores larger than 1.96 in each sample, the number of t-scores in each sample and the number of distinct research articles in each sample. Data from Brodeur et al. (2020).

1.11 Figures for Chapter 1

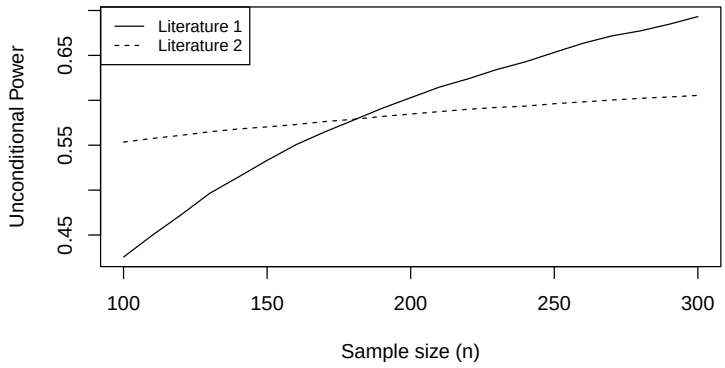

Figure 1.1. Underpowered literature vs well-powered literature.

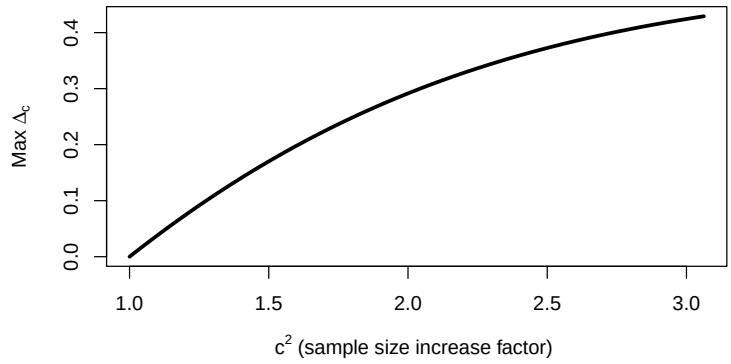

Figure 1.2. Maximum possible values of Δ_c when $cv = 1.96$.

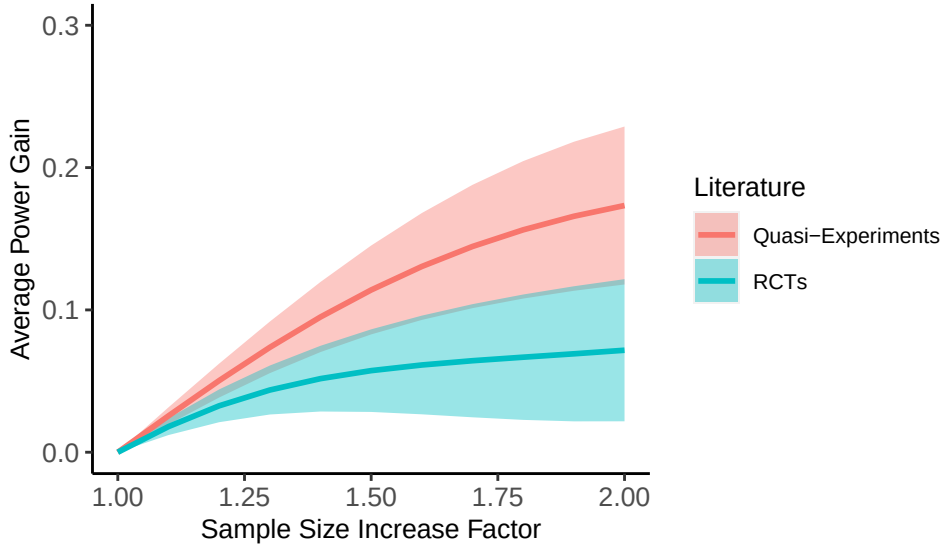

Figure 1.3. Compares power gain (y axis) of Economics RCTs vs non-RCTs over many c^2 (x-axis). Data from Brodeur et al. (2016).

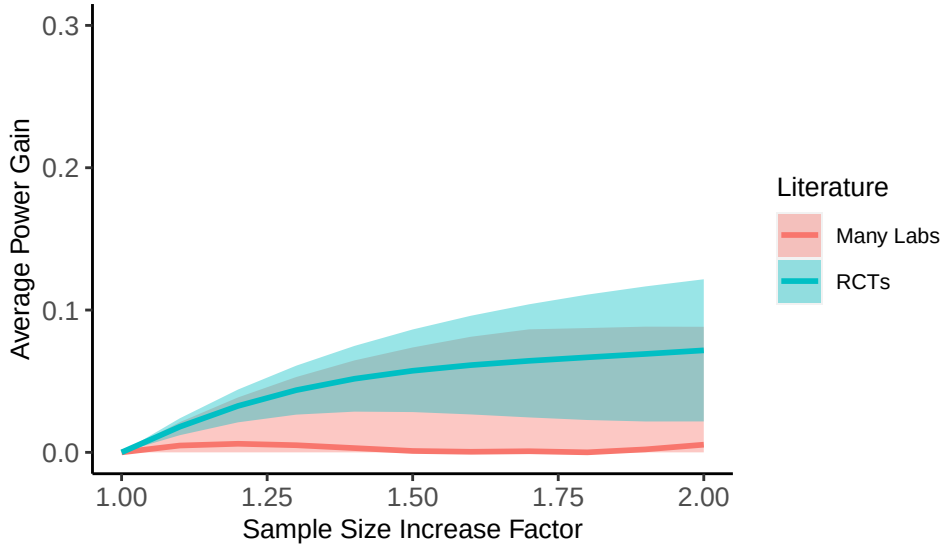

Figure 1.4. Compares power gain (y axis) of Many Labs vs Economics RCTs over many c^2 (x-axis). Data from Brodeur et al. (2016) and Klein et al. (2014).

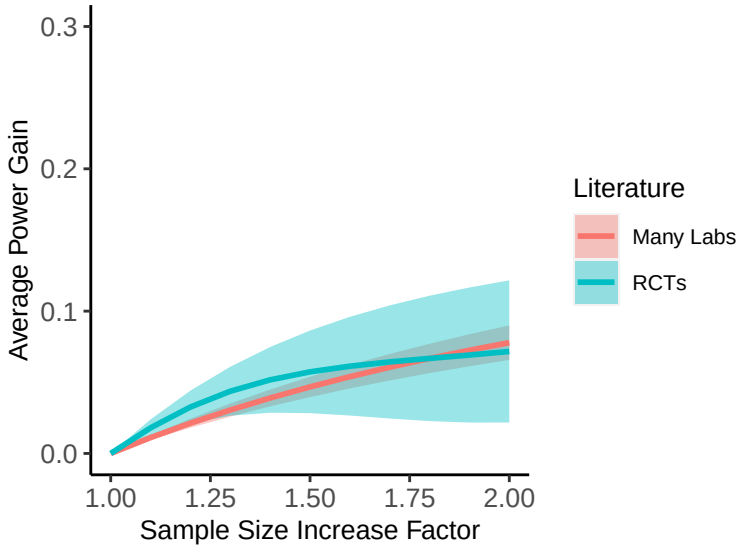

Figure 1.5. Compares power gain (y axis) of Many Labs vs Economics RCTs over many c^2 (x-axis).

1.12 Detailed Expressions

This section contains explicit expressions for terms that are too lengthy and involved for the main text.

1.12.1 Hermite Polynomials

The generalized and scaled probabalist's Hermite Polynomials are written explicitly below:

$$\chi_j(t) = \frac{1}{\sqrt{j!}} \sum_{l=0}^{[j/2]} (-1)^l \frac{(2l)!}{2^l l!} \binom{j}{2l} \left(\frac{x}{\sqrt{1+\sigma_Y^2}} \right)^{j-2l}$$

$$\phi_j(t) = \frac{1}{\sqrt{j!}} \sum_{l=0}^{[j/2]} (-1)^l \frac{(2l)!}{2^l l!} \binom{j}{2l} \left(\frac{x}{\sqrt{1+\sigma_Y^2 - c^{-2}}} \right)^{j-2l}$$

$$\psi_j(t) = \frac{1}{\sqrt{j!}} \sum_{l=0}^{[j/2]} (-1)^l \frac{(2l)!}{2^l l!} \binom{j}{2l} \left(\frac{x}{\sigma_Y} \right)^{j-2l}$$

1.12.2 Centering term $\tilde{\Delta}_{c,n}$

When we do inference we must center the estimator on a sequence $\tilde{\Delta}_{c,n}$. While $\tilde{\Delta}_{c,n}$ does converge to Δ_c, it may still be large enough to matter for inference. The full expression is below.

$$\begin{aligned}
\tilde{\Delta}_{c,n} \equiv \Delta_c &+ \sum_{j=J_n+1}^{\infty} \eta_j \langle \psi_j, \pi_0 \rangle_W a_j \\
&+ \sum_{j=0}^{J_n} a_j \frac{E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right] \left[G(\theta_0)'G(\theta_0)\right]^{-1} G(\theta_0)'E[\mathbf{v}(T) \mid R]}{\lambda_j E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \\
&+ \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0} \left[G(\theta_0)'G(\theta_0)\right]^{-1} G(\theta_0)'E[\mathbf{v}(T) \mid R]
\end{aligned}$$

where a_j is shorthand for:

$$a_j \equiv \lambda_j \int_{-cv}^{cv} \phi_j(s)ds - \int_{-cv/c}^{cv/c} \phi_j(s)ds$$

The expression references the gradients $\hat{B}$ and B which are defined as the $K \times 1$ vectors:

$$\hat{B}_\theta(t) \equiv \nabla_\theta \frac{1}{w_\theta(t)}$$
$$B_\theta \equiv E\left[\hat{B}_{\theta_0}(T) \mid R\right]$$

The $|\mathscr{X}| \times K$ population gradient matrix G is defined as:

$$G(\theta)_{x,k} = -\lim_{\varepsilon \to 0} \frac{f_{T|R}(x+\varepsilon)}{w_\theta(x+\varepsilon_n)^2} \frac{d}{d\theta} w_\theta(x+\varepsilon) + \frac{f_{T|R}(x-\varepsilon)}{w_\theta(x-\varepsilon)^2} \frac{d}{d\theta} w_\theta(x-\varepsilon)$$

$\mathbf{v}_n(t)$ is a function that maps $t \in \mathbb{R}$ to a $|\mathscr{X}| \times 1$ vector with components:

$$\mathbf{v}_{n,x}(t) = \frac{1}{\varepsilon_n} \left(\frac{\mathbf{1}\{t \in (x, x+\varepsilon_n]\}}{w_{\theta_0}(x+\varepsilon_n)} - \frac{\mathbf{1}\{t \in (x-\varepsilon_n, x]\}}{w_{\theta_0}(x-\varepsilon_n)} \right)$$

1.12.3 Linearized Functions $Z_{n,J_n}(t)$

Lemma 7 expresses estimation error in terms of a sample average of $\frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i)$. The functional form is written down below:

$$
\begin{aligned}
Z_{n,J_n}(t) = & \frac{1}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]} \sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} \frac{\psi_j(t)\varphi_{\sigma_Y}(t)}{w_{\theta_0}(t)} \\
& + \frac{\sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right] \left[G(\theta_0)'G_n(\theta_0)\right]^{-1} G(\theta_0)'\mathbf{v}(t)}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \\
& + \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \frac{1}{w_{\theta_0}(t)} \\
& + \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0} \left[G(\theta_0)'G_n(\theta_0)\right]^{-1} G(\theta_0)'\mathbf{v}(t)
\end{aligned}
$$

where a_j is shorthand for:

$$
a_j \equiv \lambda_j \int_{-cv}^{cv} \phi_j(s)ds - \int_{-cv/c}^{cv/c} \phi_j(s)ds
$$

The meta-analyst doesn't know Z_n but in Theorem 4 they use an estimate $\hat{Z}_n$ to do inference. The function $\hat{Z}_n$ is written down below:

$$\hat{Z}_{n,J_n}(t) = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}}\sum_{j=0}^{J_n}\frac{a_j}{\lambda_j}\frac{\psi_j(t)\varphi_{\sigma_Y}(t)}{w_{\hat{\theta}_n}(t)}$$
$$+\frac{\sum_{j=0}^{J_n}\frac{a_j}{\lambda_j}E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\hat{\theta}_n}(T)\mid R\right]\left[G(\hat{\theta}_n)'G_n(\hat{\theta}_n)\right]^{-1}G(\hat{\theta}_n)'\mathbf{v}_{\hat{\theta}_n}(t)}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}\right)^2}$$
$$+\frac{\hat{\Delta}_{c,n}}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}\right)^2}\frac{1}{w_{\hat{\theta}_n}(t)}$$
$$+\frac{\hat{\Delta}_{c,n}}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}\right)^2}B'_{\hat{\theta}_n}\left[G(\hat{\theta}_n)'G_n(\hat{\theta}_n)\right]^{-1}G(\hat{\theta}_n)'\mathbf{v}_{\hat{\theta}_n}(t)$$

1.13 Proofs for Chapter 1

1.13.1 Proof of Lemma 1

We show this using characteristic functions. $K_1\pi$ is the PDF of the random variable $h+Z$ where $Z\sim N(0,1)$ and Z is independent of h. By the Convolution Theorem, the characteristic function of the sum of any two independent random variables is the pointwise product of the characteristic functions of the summands. The characteristic function of a normal random variable with mean zero and variance σ^2 is $e^{\omega^2/2\sigma^2}$. Let $\zeta_h(\omega)$ denote the characteristic function of h. The characteristic function of $h+Z$ is $e^{-\omega^2/2}\zeta_h$.

Now consider $K_{1-c^{-2}}K_{c^{-2}}\pi$. This is the PDF of the random variable $h+c^{-1}Z_1+\sqrt{1-c^{-2}}Z_2$ where h,Z_1,Z_2 are all independent and $Z_1,Z_2\sim N(0,1)$. To find the characteristic function of the sum, we take the pointwise product of the chacteristic functions of the summands which is: $e^{-\omega^2/2c^{-2}}e^{-\omega^2/2(1-c^{-2})}\zeta_h = e^{-\omega^2/2}\zeta_h$. So $K_1\pi$ and $K_{1-c^{-2}}K_{c^{-2}}\pi$ are PDFs that share the same characteristic function. So they must agree almost everywhere and be the same distribution.

1.13.2 Proof of Lemma 2

It is easier to first decompose Δ_c into the difference of the type-II error probabilities:

$$\begin{aligned}
\Delta_c &= \beta_1 - \beta_c \\
\beta_c &= \int_{-cv}^{cv} f_{T_c}(t)dt \\
\beta_1 &= \int_{-cv}^{cv} f_T(t)dt
\end{aligned}$$

We begin with the definition of β_c. Then we do a change of variables of integration. Then we substitute in the definition of $K_{c^{-2}}$.

$$\begin{aligned}
\beta_c &= \int_{-cv}^{cv} f_{T_c}(t)dt \\
&= \int_{-cv}^{cv}\int_{-\infty}^{\infty} \varphi(t-ch)\pi_0(h)dhdt \\
&= \int_{-cv/c}^{cv/c}\int_{-\infty}^{\infty} \varphi(ct-ch)\pi_0(h)dhdt \\
&= \int_{-cv/c}^{cv/c}\int_{-\infty}^{\infty} \varphi_{c^{-2}}(t-h)\pi_0(h)dhdt \\
&= \int_{-cv/c}^{cv/c} K_{c^{-2}}\pi_0(h)dt
\end{aligned}$$

By an identical argument:

$$\beta_1 = \int_{-cv}^{cv} K_1\pi_0(h)dt$$

Taking the difference:

$$\Delta_c = \int_{-cv}^{cv} (K_{1-c^{-2}}K_{c^{-2}}\pi_0)[t]dt - \int_{-cv/c}^{cv/c} (K_{c^{-2}}\pi_0)[t]dt$$

1.13.3 Proof of Lemma 3

As in Lemma 2, is easier to first decompose Δ_c into the difference of the type-II error probabilities:

$$\begin{aligned}
\Delta_c &= \beta_1 - \beta_c \\
\beta_c &= \int_{-cv}^{cv} f_{T_c}(t)dt \\
\beta_1 &= \int_{-cv}^{cv} f_T(t)dt
\end{aligned}$$

Using the same argument as in Lemma 2, we do a change of variables of integration. Then we substitute in the definition of $K_{c^{-2}}$. Then we substitute in the singular value decomposition from (Carrasco and Florens, 2011).

$$\begin{aligned}
\beta_c &= \int_{-cv}^{cv} f_{T_c}(t)dt \\
&= \int_{-cv}^{cv}\int_{-\infty}^{\infty} \varphi(t-ch)\pi_0(h)dhdt \\
&= \int_{-cv/c}^{cv/c}\int_{-\infty}^{\infty} \varphi(ct-ch)\pi_0(h)dhdt \\
&= \int_{-cv/c}^{cv/c}\int_{-\infty}^{\infty} \varphi_{c^{-2}}(t-h)\pi_0(h)dhdt \\
&= \int_{-cv/c}^{cv/c} K_{c^{-2}}\pi_0 dt \\
&= \int_{-cv/c}^{cv/c} \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_0\rangle_W \phi_j(t)dt
\end{aligned}$$

We would like to exchange the sum and the integral with Fubini's Theorem. To do this it is sufficient to show that:

$$\sum_{j=0}^{\infty} \eta_j \left|\langle \chi_j, \pi_0\rangle_W\right| \int_{-cv/c}^{cv/c} \left|\phi_j(t)\right| dt < \infty$$

By the Cauchy Schwarz Inequality:

$$\begin{aligned}\int_{-cv/c}^{cv/c} \phi_j(t)dt &\leq \int_{-cv/c}^{cv/c} |\phi_j(t)|dt \\ &\leq 2\frac{cv}{c}\int_{-cv/c}^{cv/c} \phi_j(t)^2 dt\end{aligned}$$

The singular functions are normalized in their respective Hilbert spaces so $||\phi_j||_X = 1$. This is the same as taking the inner product with itself: $||\phi_j||_X \equiv \langle \phi_j, \phi_j \rangle_X = \int_{-\infty}^{\infty} \phi_j(t)^2 \varphi_{1+\sigma_Y^2+c^{-2}}(t)dt = 1$. Since the integrand is non-negative, this implies that bounded integral is less than one. $\int_{-cv/c}^{cv/c} \phi_j(t)^2 \varphi_{1+\sigma_Y^2+c^{-2}}(t)dt \leq 1$. Now we use the fact that the normal density is bounded below on this compact set:

$$\min_{t \in -cv/c, cv/c} \varphi_{1+\sigma_Y^2+c^{-2}}(t) = \varphi_{1+\sigma_Y^2+c^{-2}}(cv/c) = \frac{e^{\frac{-cv^2}{c^2(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+c^{-2})}}$$

Since the minimum of $\varphi_{1+\sigma_Y^2+c^{-2}}(t)$ is bounded below, the integral without the normal kernel can be bounded above:

$$\int_{-cv/c}^{cv/c} \phi_j(t)^2 dt \leq \frac{e^{\frac{cv^2}{c^2(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+c^{-2})}}$$

Combining this with the Cauchy Schwarz bound from before:

$$\int_{-cv/c}^{cv/c} |\phi_j(t)|dt \leq 2\frac{cv}{c}\frac{e^{\frac{cv^2}{c^2(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+c^{-2})}}$$

This bound applies uniformly to all of the integrals for all j. This gives us the

following bound on the absolute sum of the absolute values.

$$\sum_{j=0}^{\infty} \eta_j \left|\langle \chi_j, \pi_0 \rangle_W\right| \int_{-cv/c}^{cv/c} \left|\phi_j(t)\right| dt \leq \left(2\frac{cv}{c} \frac{e^{\frac{cv^2}{c^2(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+c^{-2})}} \right) \sum_{j=0}^{\infty} \eta_j \left|\langle \chi_j, \pi_0 \rangle_W\right|$$

Since π_0 is a PDF: $|\langle \chi_j, \pi_0 \rangle_W| \leq 1$ by Holder's Inequality for all j. Since the η_j are a positive power series, their sum converges. So we have proven that:

$$\sum_{j=0}^{\infty} \eta_j \left|\langle \chi_j, \pi_0 \rangle_W\right| \int_{-cv/c}^{cv/c} \left|\phi_j(t)\right| dt < \infty$$

Since the absolute sum of the absolute integrals is finite we can use Fubini's Theorem and exchange the sum and the integral to obtain the β_c as a sum.

$$\beta_c = \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_0 \rangle_W \int_{-cv/c}^{cv/c} \phi_j(t) dt$$

An identical argument with $c = 1$ yields:

$$\beta_1 = \sum_{j=0}^{\infty} \eta_j \lambda_j \langle \chi_j, \pi_0 \rangle_W \int_{-cv}^{cv} \phi_j(t) dt$$

We can use this formula to convert back into Δ_c:

$$\Delta_c = \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_0 \rangle_W \left(\lambda_j \int_{-cv}^{cv} \phi_j(t) dt - \int_{-cv/c}^{cv/c} \phi_j(t) dt \right)$$

1.13.4 Proof of Proposition 1

This proof formalizes the argument in Section 1.4.1. We start with the definition of Q_0. We then rewrite this in terms of the inner product.

$$\begin{aligned}
Q_0(\pi) &\equiv \int_{-\infty}^{\infty} \left(f_T(t) - \left(\int_{-\infty}^{\infty} \varphi(t-h)\pi(h)dh \right) \right)^2 \varphi_{\sigma_Y^2}(t)dt \\
&= \int_{-\infty}^{\infty} \left((K_1\pi_0)[t] - (K_1\pi)[t] \right)^2 \varphi_{\sigma_Y^2}(t)dt \\
&= \langle K_1\pi_0 - K_1\pi, K_1\pi_0 - K_1\pi \rangle_W \\
&\equiv ||K_1\pi_0 - K_1\pi||_Y \\
&= ||K_1(\pi_0 - \pi)||_Y
\end{aligned}$$

Next we substitue in $K_1 = K_{1-c^{-1}} K_{c^{-2}}$ from Lemma 1:

$$||K_1(\pi_0 - \pi)||_Y = \left|\left|K_{1-c^{-1}} K_{c^{-2}}(\pi_0 - \pi)\right|\right|_Y$$

Recall the singular value decompositions of $K_{1-c^{-1}}$ and $K_{c^{-2}}$ from Section 1.3.3:

$$\begin{aligned}
K_{c^{-2}}\pi &= \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi \rangle_W \phi_j \\
K_{1-c^{-2}} g &= \sum_{j=0}^{\infty} \lambda_j \langle \phi_j, g \rangle_X \psi_j
\end{aligned}$$

Substituting in the expressions for the singular value decomposition and then

using the fact that the ϕ_j are orthonormal in $\langle \cdot, \cdot \rangle_X$:

$$\begin{aligned} K_{1-c^{-1}} K_{c^{-2}} \pi &= \sum_{j=0}^{\infty} \lambda_j \left\langle \phi_j, K_{c^{-2}} \pi \right\rangle_X \psi_j \\ &= \sum_{j=0}^{\infty} \lambda_j \left\langle \phi_j, \sum_{k=0}^{\infty} \eta_j \langle \chi_k, \pi \rangle_W \phi_k \right\rangle_X \psi_j \\ &= \sum_{j=0}^{\infty} \lambda_j \eta_j \langle \chi_k, \pi \rangle_W \psi_j \end{aligned}$$

Substituting the SVD into the objective function:

$$\left|\left| K_{1-c^{-1}} K_{c^{-2}} (\pi_0 - \pi) \right|\right|_Y = \left|\left| \sum_{j=0}^{\infty} \lambda_j \eta_j \langle \chi_k, \pi_0 - \pi \rangle_W \psi_j \right|\right|_Y$$

Since ψ_j are orthogonal in $\langle \cdot, \cdot \rangle_Y$ we can use the Pythagorean Formula to write the norm as the sum of inner products. Then we can use the orthonormality of ψ_j a second time to eliminate the cross terms.

$$\begin{aligned} \left|\left| \sum_{j=0}^{\infty} \lambda_j \eta_j \langle \chi_k, \pi_0 - \pi \rangle_W \psi_j \right|\right|_Y &= \sum_{j=0}^{\infty} \left\langle \sum_{k=0}^{\infty} \lambda_k \eta_k \langle \chi_j, \pi_0 - \pi \rangle_W \psi_k, \psi_j \right\rangle_Y^2 \\ &= \sum_{j=0}^{\infty} \left(\sum_{k=0}^{\infty} \left\langle \lambda_k \eta_k \langle \chi_j, \pi_0 - \pi \rangle_W \psi_k, \psi_j \right\rangle_Y \right)^2 \\ &= \sum_{j=0}^{\infty} \lambda_j^2 \eta_j^2 \left\langle \chi_j, \pi_0 - \pi \right\rangle_W^2 \end{aligned}$$

The singular values η_j, λ_j are all strictly positive. Since the generalized Hermite polynomials χ_j form a complete basis, any function in $\mathscr{L}_W$ that is orthogonal to all of the χ_j must be equal zero almost everywhere (Johnston, 2014). This means that any π that satisfies $Q_0(\pi) = 0$ must equal π_0 almost everywhere. If two probability density functions agree almost everywhere, then they are the same density. So $Q_0(\pi) = 0$ if and only if $\pi_0 = 0$. Since Q_0 is everywhere non-negative this gives us identification of π_0.

1.13.5 Proof of Lemma 4

First we decompose Δ_c into the difference between type-II error probabilities as we did in the proof of Lemma 3:

$$\begin{aligned}\Delta_c &= \beta_1 - \beta_c \\ \beta_c &= \int_{-cv}^{cv} f_{T_c}(t)dt \\ \beta_1 &= \int_{-cv}^{cv} f_T(t)dt\end{aligned}$$

In the proof of Lemma 3 we showed that:

$$\begin{aligned}\beta_c &= \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_0 \rangle_W \int_{-cv/c}^{cv/c} \phi_j(t)dt \\ \beta_1 &= \sum_{j=0}^{\infty} \eta_j \lambda_j \langle \chi_j, \pi_0 \rangle_W \int_{-cv}^{cv} \phi_j(t)dt\end{aligned}$$

In the proof of Lemma 3 we also showed that:

$$\int_{-cv/c}^{cv/c} |\phi_j(t)|dt \leq 2\frac{cv}{c} \frac{e^{\frac{cv^2}{c^2(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+c^{-2})}}$$

This means that for any $\pi \in \mathscr{L}_W$,

$$\begin{aligned}
\left|\beta_c - \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi \rangle_W \int_{-cv/c}^{cv/c} \phi_j(t)dt\right| &= \left|\sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_0 - \pi \rangle_W \int_{-cv/c}^{cv/c} \phi_j(t)dt\right| \\
&\leq \sum_{j=0}^{\infty} \eta_j \left|\langle \chi_j, \pi_0 - \pi \rangle_W\right| \left|\int_{-cv/c}^{cv/c} \phi_j(t)dt\right| \\
&\leq \left(2\frac{cv}{c} \frac{e^{\frac{cv^2}{c^2(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+c^{-2})}}\right) \sum_{j=0}^{\infty} \eta_j \left|\langle \chi_j, \pi_0 - \pi \rangle_W\right| \\
&= \left(2\frac{cv}{c} \frac{e^{\frac{cv^2}{c^2(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+c^{-2})}}\right) \rho_c(\pi_0, \pi)
\end{aligned}$$

By an identical argument:

$$\left|\beta_1 - \sum_{j=0}^{\infty} \eta_j \lambda_j \langle \chi_j, \pi \rangle_W \int_{-cv}^{cv} \phi_j(t)dt\right| \leq \left(2cv \frac{e^{\frac{cv^2}{(1+\sigma_Y^2+c^{-2})}}}{\sqrt{2\pi(1+\sigma_Y^2+1)}}\right) \rho_c(\pi_0, \pi)$$

Since the constant in front does not depend on π, for any sequence of π_n

$$\rho_c(\pi_0, \pi_n) = O_p(r_b) \implies \Delta_c - \sum_{j=0}^{\infty} \eta_j \langle \chi_j, \pi_n \rangle_W \left(\lambda_j \int_{-cv}^{cv} \phi_j(t)dt - \int_{-cv/c}^{cv/c} \phi_j(t)dt\right) = O_p(r_n)$$

1.13.6 Proof of Theorem 1

From definition of ρ_c:

$$\rho_c(\pi_0, \hat{\pi}_n) = \sum_{j=0}^{\infty} \eta_j \left|\langle \chi_j, \pi_0 \rangle_W - \langle \chi_j, \hat{\pi}_n \rangle_W\right|$$

The spectral cutoff regularization of $\hat{\pi}_n$ sets: $\langle \chi_j, \hat{\pi}_n \rangle_W = 0 \, \forall_{j > J_n}$. So we can split

ρ_c into the sampling error and regularization bias parts:

$$\rho_c(\pi_0, \hat{\pi}_n) = \sum_{j=0}^{J_n} \eta_j \left| \langle \chi_j, \pi_0 \rangle_W - \langle \chi_j, \hat{\pi}_n \rangle_W \right| + \sum_{J_n+1}^{\infty} \eta_j \left| \langle \chi_j, \pi_0 \rangle_W \right|$$

Now we bound the first sum which comes from sampling error. The first step is to substitute in $\hat{\pi}_n$ from Equation 1.13. Then we use the linearity of the inner product and the fact that the χ_j are orthonormal in $\langle \cdot, \cdot \rangle_W$.

$$\begin{aligned}
\sum_{j=0}^{J_n} \eta_j \left| \langle \chi_j, \pi_0 \rangle_W - \langle \chi_j, \hat{\pi}_n \rangle_W \right| &= \sum_{j=0}^{J_n} \eta_j \left| \langle \chi_j, \pi_0 \rangle_W - \left\langle \chi_j, \sum_{k=0}^{J_n} \frac{\frac{1}{n}\sum_{i=1}^n \psi_k(t_i)\varphi_{\sigma_Y^2}(t_i)}{\eta_k \lambda_k} \chi_k \right\rangle_W \right| \\
&= \sum_{j=0}^{J_n} \eta_j \left| \langle \chi_j, \pi_0 \rangle_W - \sum_{k=0}^{J_n} \left\langle \chi_j, \frac{\frac{1}{n}\sum_{i=1}^n \psi_k(t_i)\varphi_{\sigma_Y^2}(t_i)}{\eta_k \lambda_k} \chi_k \right\rangle_W \right| \\
&= \sum_{j=0}^{J_n} \eta_j \left| \langle \chi_j, \pi_0 \rangle_W - \frac{\frac{1}{n}\sum_{i=1}^n \psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{\eta_j \lambda_j} \right| \\
&= \sum_{j=0}^{J_n} \left| \eta_j \langle \chi_j, \pi_0 \rangle_W - \frac{\frac{1}{n}\sum_{i=1}^n \psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{\lambda_j} \right|
\end{aligned}$$

We now wish to substitute in an expectation over T for $\eta_j \langle \chi_j, \pi_0 \rangle_W$. We can do this with equation 1.6 which says:

$$f_T = \sum_{j=1}^{\infty} \lambda_j \eta_j \langle \chi_j, \pi_0 \rangle_W \psi_j$$

Since the ψ_j are orthonormal:

$$\langle f_T, \psi_j \rangle_Y = \lambda_j \eta_j \langle \chi_j, \pi_0 \rangle_W$$

The inner $\langle \cdot, \cdot \rangle_Y$ product is an integral and f_T is a PDF. The integral over a PDF is

an expectation. So we have:

$$E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right] = \lambda_j \eta_j \langle \chi_j, \pi_0 \rangle_W$$
$$\frac{E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right]}{\lambda_j} = \eta_j \langle \chi_j, \pi_0 \rangle_W$$

Substituting this into our expression for the sampling error:

$$\sum_{j=0}^{J_n} \left| \eta_j \langle \chi_j, \pi_0 \rangle_W - \frac{\frac{1}{n}\sum_{i=1}^n \psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{\lambda_j} \right| = \sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right] - \frac{1}{n}\sum_{i=1}^n \psi_j(t_i)\varphi_{\sigma_Y^2}(t_i) \right|$$

The observations t_i are identically distributed draws of T. While they are not all independent, each observation is independent of at least $n - C$ other observations. So, the variance is bounded by:

$$\begin{aligned} Var\left(\frac{1}{n}\sum_{i=1}^n \psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)\right) &= \frac{1}{n^2}\sum_{i=1}^n\sum_{l=1}^n Cov\left(\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i), \psi_j(t_l)\varphi_{\sigma_Y^2}(t_l)\right) \\ &\leq C\frac{1}{n^2}\sum_{i=1}^n\sum_{l=1}^n Var\left(\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)\right) \\ &= \frac{C}{n} Var\left(\psi_j(T)\varphi_{\sigma_Y^2}(T)\right) \end{aligned}$$

The distribution of T is not changing. Furthermore, we upper bound $Var\left(\psi_j(T)\varphi_{\sigma_Y^2}(T)\right)$ by 1 using Holder's Inequality and the fact that $||\psi||_Y = 1$.

$$\begin{aligned} Var\left(\psi_j(T)\varphi_{\sigma_Y^2}(T)\right) &\leq E\left[\left(\psi_j(T)\varphi_{\sigma_Y^2}(T)\right)^2\right] \\ &= \int_{-\infty}^{\infty} \psi_j(t)^2 \varphi_{\sigma_Y^2}(t)^2 f_T(t) dt \\ &\leq \int_{-\infty}^{\infty} \psi_j(t)^2 \varphi_{\sigma_Y^2}(t) dt \\ &= \langle \psi_j, \psi_j \rangle_Y \\ &= 1 \end{aligned}$$

This lets us upper bound the variance of the sums:

$$Var\left(\frac{1}{n}\sum_{i=1}^{n}\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)\right) \leq \frac{C}{n}$$

By Chebyshev's Inequality:

$$\frac{1}{n}\sum_{i=1}^{n}\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i) = O_p(n^{-1/2})$$

Now we want to bound the sum of the centered sample means over j in the sampling error. use the fact that λ_j is a power sequence, i.e.

$$\lambda_j = \lambda_0^{j/2}$$

This gives us the identity:

$$\begin{aligned}
\lambda_{J_n}\sum_{j=0}^{J_n}\frac{1}{\lambda_j} &= \sum_{j=0}^{J_n}\lambda_0^{J_n - j/2} \\
&= \sum_{j=0}^{J_n}\lambda_0^{j/2} \\
&= \sum_{j=0}^{J_n}\lambda_j \\
&\to \frac{1}{1-\lambda_0}
\end{aligned}$$

This means that we can control the order of the sum.

$$\sum_{j=0}^{J_n}\frac{1}{\lambda_j} = O\left(\lambda_{J_n}^{-1}\right)$$

Combining this with the order of the centered sample means we can control the

order of the sampling error part:

$$\sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right] - \frac{1}{n}\sum_{i=1}^{n} \psi_j(t_i)\varphi_{\sigma_Y^2}(t_i) \right| = O_p\left(\lambda_{J_n}^{-1} n^{-1/2}\right)$$

Next we bound the second sum that comes from regularization bias. Since π_0 is a PDF, its $\mathscr{L}_1$ norm is 1. Since the normal density with variance greater than 1 is always itself less than 1 the $||\cdot||_Y$-norm is less than the $\mathscr{L}_1$ norm and $||\pi_0||_W \leq 1$. Using the Cauchy Schwarz Indequality and the fact that the $\eta_j = \eta_2^{j/2}$:

$$\begin{aligned}
\sum_{J_n+1}^{\infty} \eta_j \left|\langle \chi_j, \pi_0 \rangle_W\right| &\leq \sqrt{\sum_{J_n+1}^{\infty} \eta_j^2} \sqrt{\sum_{J_n+1}^{\infty} \langle \chi_j, \pi_0 \rangle_W^2} \\
&\leq \sqrt{\sum_{J_n+1}^{\infty} \eta_j^2} \\
&= \eta_{J_n+1} \sqrt{\sum_{J_n+1}^{\infty} \frac{\eta_j^2}{\eta_{J_n+1}^2}} \\
&= \eta_{J_n+1} \sqrt{\sum_{J_n+1}^{\infty} \left(\frac{\eta_2^{j/2}}{\eta_2^{(J_n+1)/2}}\right)} \\
&= \eta_{J_n+1} \sqrt{\sum_{0}^{\infty} \eta_2^{j/2}} \\
&= \eta_{J_n+1} \frac{1}{1-\eta_2} \\
&\leq \frac{\eta_{J_n+1}}{1-\eta_2} \\
&= O\left(\eta_{J_n}\right)
\end{aligned}$$

Combining the bounds on the sampling error and regularization bias we have proven the first claim in Theorem 1:

$$\rho_c(\pi_0, \hat{\pi}_n) = O_p\left(\lambda_{J_n}^{-1} n^{-1/2} + \eta_{J_n}\right)$$

Now we prove the second claim in Theorem 1. Our goal is to choose a growth rate for J_n given $\sigma_Y^2 > 0$ such that the order $\lambda_{J_n}^{-1} n^{-1/2} + \eta_{J_n}$ is minimized. To do this it is sufficient to set the orders of the two summands equal to each other. A sufficient condition for this is that for some $d > 0$,

$$\frac{\eta_{J_n}}{\lambda_{J_n}^{-1} n^{-1/2}} \to c$$

Recall the definitions of the singular values λ_j and η_j from Section 1.3.3.

$$\eta_j = \left(\frac{1 + \sigma_Y^2 - c^{-2}}{1 + \sigma_Y^2} \right)^{j/2}$$
$$\lambda_j = \left(\frac{\sigma_Y^2}{\sigma_Y^2 + 1 - c^{-2}} \right)^{j/2}$$

Notice that we can rewrite η_j as:

$$\eta_j = \lambda_j^{-1} \left(\frac{\sigma_Y^2}{1 + \sigma_Y^2} \right)^{j/2}$$

Which makes the ratio:

$$\frac{\eta_{J_n}}{\lambda_{J_n}^{-1} n^{-1/2}} = \left(\frac{\sigma_Y^2}{1 + \sigma_Y^2} \right)^{j/2} n^{1/2}$$

So if the meta-analyst chooses J_n such that for some $c > 0$, $n^{1/2} \left(\frac{\sigma_Y^2}{\sigma_Y^2 + 1} \right)^{J_n/2} \to c$, then:

$$\frac{\eta_{J_n}}{\lambda_{J_n}^{-1} n^{-1/2}} \to c$$

So the order of η_{J_n} and $n^{-1/2} \lambda_{J_n}^{-1}$ are the same and the order of the sum is

minimized. Next we compute this order exactly. To do this we compute:

$$\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)^{J_n/2} = O\left(n^{-1/2}\right)$$

$$J_n/2 = O\left(\log_{\frac{\sigma_Y^2}{\sigma_Y^2+1}}\left(n^{-1/2}\right)\right)$$

$$\eta_{J_n} = O\left(\left(\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}\right)^{\log_{\frac{\sigma_Y^2}{\sigma_Y^2+1}}\left(n^{-1/2}\right)}\right)$$

Using the change of log base:

$$\log_{\frac{\sigma_Y^2}{\sigma_Y^2+1}}\left(n^{-1/2}\right) = \frac{\log_{\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}}\left(n^{-1/2}\right)}{\log_{\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}}\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)}$$

This lets us rewrite:

$$\begin{aligned}
\left(\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}\right)^{\log_{\frac{\sigma_Y^2}{\sigma_Y^2+1}}\left(n^{-1/2}\right)} &= \left(\left(\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}\right)^{\log_{\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}}\left(n^{-1/2}\right)}\right)^{1/\log_{\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}}\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)} \\
&= \left(n^{-1/2}\right)^{1/\log_{\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}}\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)} \\
&= n^{-1/2\log_{\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}}\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)} \\
&= n^{-\frac{1}{2}\frac{\log\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)}{\log\left(\frac{1+\sigma_Y^2-c^{-2}}{1+\sigma_Y^2}\right)}} \\
&= n^{-\frac{1}{2}\frac{\log\left(\frac{1+\sigma_Y^2}{1+\sigma_Y^2-c^{-2}}\right)}{\log\left(\frac{\sigma_Y^2+1}{\sigma_Y^2}\right)}} \\
&= n^{-q/2}
\end{aligned}$$

As claimed in Theorem 1.

1.13.7 Proof of Lemma 5

By Bayes' Rule:

$$f_T(t|R) = \frac{f_T(t)w_{\theta_0}(t)}{E[w_{\theta_0}(T)]}$$
$$f_T(t) = \frac{f_T(t|R)E[w_{\theta_0}(T)]}{w_{\theta_0}(t)}$$

For any measurable function $\gamma : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned} E[\gamma(T)] &= \int_{-\infty}^{\infty} \gamma(t) f_T(t) dt \\ &= \int_{-\infty}^{\infty} \gamma(t) \frac{f_T(t)|R}{w_{\theta_0}(t)} dt E[w_{\theta_0}(T)] \\ &= E\left[\frac{\gamma(T)}{w_{\theta_0}(T)} \,|\, R\right] E[w_{\theta_0}(T)] \end{aligned}$$

To compute $E[w_{\theta_0}(T)]$ by taking an expectation of over $T|R$, we use the following trick:

$$\begin{aligned} E\left[\frac{1}{w_{\theta_0}(T)}|R\right] &= \int_{-\infty}^{\infty} \frac{1}{w_{\theta_0}(T)} f_T(t|R) dt \\ &= \int_{-\infty}^{\infty} \frac{1}{w_{\theta_0}(T)} \frac{f_T(t)w_{\theta_0}(t)}{E[w_{\theta_0}(T)]} dt \\ &= \frac{1}{E[w_{\theta_0}(T)]} \end{aligned}$$

Putting the two pieces together:

$$E[\gamma(T)] = \frac{E\left[\frac{\gamma(T)}{w_{\theta_0}(T)} \,|\, R\right]}{E[w_{\theta_0}(T) \,|\, R]}$$

1.13.8 Proof of Proposition 2

Recall the population objective function:

$$Q_0(\pi,\theta) \equiv \sum_{j=0}^{\infty}\left(E\left[\frac{\psi_j(T)\,\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)}\,|\,R\right]/E\left[\frac{1}{w_{\theta_0}(T)}\,|\,R\right]-\lambda_j\eta_j\langle\pi,\chi_j\rangle_W\right)^2 + \sum_{x\in\mathscr{X}}\lim_{\varepsilon\to 0}\left(\frac{f_{T\,|\,R}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{f_{T\,|\,R}(x-\varepsilon)}{w_\theta(x-\varepsilon)}\right)^2$$

As in the case without publication bias, $Q_0(\pi_0,\theta_0)=0$ and $Q_0(\pi,\theta)$ is everywhere non-negative. To show identification we must show that the objective cannot equal zero for any other arguments. First we show that if $\theta\neq\theta_0$ then $Q_0(\pi,\theta_0)>0$ for any π. Using Bayes' Rule:

$$f_T(t|R)=\frac{f_T(t)w_{\theta_0}(t)}{E[w_{\theta_0}(T)]}$$

Since f_T must be continuous,

$$\begin{aligned}
&\sum_{x\in\mathscr{X}}\lim_{\varepsilon\to 0}\left(\frac{f_{T\,|\,R}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{f_{T\,|\,R}(x-\varepsilon)}{w_\theta(x-\varepsilon)}\right)^2\\
&=\frac{1}{E[w_{\theta_0}(T)]}\sum_{x\in\mathscr{X}}\lim_{\varepsilon\to 0}\left(\frac{f_T(x+\varepsilon)w_{\theta_0}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{f_T(x-\varepsilon)w_{\theta_0}(x+\varepsilon)}{w_\theta(x-\varepsilon)}\right)^2\\
&=\frac{1}{E[w_{\theta_0}(T)]}\sum_{x\in\mathscr{X}}\lim_{\varepsilon\to 0}f_T(x)\left(\frac{w_{\theta_0}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{w_{\theta_0}(x+\varepsilon)}{w_\theta(x-\varepsilon)}\right)^2\\
&\geq\frac{1}{E[w_{\theta_0}(T)]}\max_{x\in\mathscr{X}}\lim_{\varepsilon\to 0}f_T(x)\left(\frac{w_{\theta_0}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{w_{\theta_0}(x+\varepsilon)}{w_\theta(x-\varepsilon)}\right)^2
\end{aligned}$$

Now recall the inequality form Assumption 3:

$$\max_{x\in\mathscr{X}}\lim_{\varepsilon\to 0}L\left|\frac{w_{\theta_0}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{w_{\theta_0}(x-\varepsilon)}{w_\theta(x-\varepsilon)}\right|\geq||\theta-\theta_0||_\infty$$

Substituting this in:

$$\frac{1}{E[w_{\theta_0}(T)]}\max_{x\in\mathscr{X}}\lim_{\varepsilon\to 0} f_T(x)\left(\frac{w_{\theta_0}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{w_{\theta_0}(x+\varepsilon)}{w_\theta(x-\varepsilon)}\right)^2 \geq \frac{1}{E[w_{\theta_0}(T)]}||\theta-\theta_0||_\infty^2 \min_{x\in\mathscr{X}} f_T(x)$$

Since $\mathscr{X}$ is finite and f_T is supported everywhere, $\min_{x\in\mathscr{X}} f_T(x)\geq 0$. Moreover, Assumption 2 guarantees that $0<\frac{1}{E[w_{\theta_0}(T)]}<\infty$. So, the second sum in the objective function is zero if and only if $\theta=\theta_0$.

$$\sum_{x\in\mathscr{X}}\lim_{\varepsilon\to 0}\left(\frac{f_{T|R}(x+\varepsilon)}{w_\theta(x+\varepsilon)}-\frac{f_{T|R}(x-\varepsilon)}{w_\theta(x-\varepsilon)}\right)^2=0 \iff ||\theta-\theta_0||_\infty=0$$

This means that in order for $Q_0(\pi,\theta)=0$, it must be $Q(\pi,\theta_0)=0$. Concentrating the objective function gives:

$$Q_0(\pi,\theta_0)=\sum_{j=0}^{\infty}\left(E\left[\frac{\psi_j(T)\,\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)}\,|\,R\right]/E\left[\frac{1}{w_{\theta_0}(T)}\,|\,R\right]-\lambda_j\eta_j\langle\pi,\chi_j\rangle_W\right)^2$$

Now we use Lemma 5 to substitute in the unconditional expectations:

$$Q_0(\pi,\theta_0)=\sum_{j=0}^{\infty}\left(E\left[\psi_j(T)\,\varphi_{\sigma_Y}(T)\right]-\lambda_j\eta_j\langle\pi,\chi_j\rangle_W\right)^2$$

By Proposition 1, this is uniquely minimized at π_0. So π_0,θ_0 are the unique minimizers of $Q_0(\pi,\theta)$ and are therefore identified.

1.13.9 Proof of Theorem 2

This proof consistes of two steps. In the first step we show the consistency of $\hat{\theta}_n$. In the second step we use this find the rate of convergence of $\rho_c(\pi_0,\hat{\pi}_n^{pb})$.

Step 1

$\hat{\theta}_n$ is the minimizer of the following sample objective function (which is $\hat{Q}_n(\pi,\theta)$

with the π concentrated out).

$$\begin{aligned}\hat{\theta}_n &= \min_{\theta} \hat{Q}_n(\hat{\pi}_n, \theta) \\ &= \min_{\theta} \sum_{x \in \mathscr{X}} \left(\frac{\frac{1}{n\varepsilon_n} \sum_{i=1}^n \mathbf{1}\{t_i \in (x, x+\varepsilon_n]\}}{w_\theta(x+\varepsilon_n)} - \frac{\frac{1}{n\varepsilon_n} \sum_{i=1}^n \mathbf{1}\{t_i \in (x-\varepsilon_n, x]\}}{w_\theta(x-\varepsilon_n)} \right)^2\end{aligned}$$

This is the analogue of the concentrated population objective function.

$$\begin{aligned}\theta_0 &= \min_{\theta} Q_0(\pi_0, \theta) \\ &= \min_{\theta} \sum_{x \in \mathscr{X}} \left(\frac{\frac{1}{n\varepsilon_n} \sum_{i=1}^n \mathbf{1}\{t_i \in (x, x+\varepsilon_n]\}}{w_\theta(x+\varepsilon_n)} - \frac{\frac{1}{n\varepsilon_n} \sum_{i=1}^n \mathbf{1}\{t_i \in (x-\varepsilon_n, x]\}}{w_\theta(x-\varepsilon_n)} \right)^2\end{aligned}$$

Next we show that the objective functions converge.

$n^{2/3} \sup_\theta \left| \hat{Q}_n(\hat{\pi}_n, \theta) - Q_0(\pi_0, \theta) \right| = O_p(1)$. We will call this fact (i).

By Assumption 3, $||\theta_0 - \hat{\theta}_n||_\infty^2 \leq L Q_0(\pi_0, \hat{\theta}_n)$. Call this fact (ii) Moreover, it is clear by inspection that $Q_0(\pi_0, \theta_0) = 0$. Call this fact (iii).

Now we will successively apply facts (i), (iii), the definition of $\hat{\theta}_n$ as the minimizer, (iii), and (ii):

$$\begin{aligned}n^{2/3} ||\theta_0 - \hat{\theta}_n||_\infty^2 &\leq n^{2/3} L Q_0(\pi_0, \hat{\theta}_n) \\ &= n^{2/3} L \hat{Q}_n(\pi_0, \hat{\theta}_n) + O_p(1) \\ &\leq n^{2/3} L \hat{Q}_n(\pi_0, \theta_0) + O_p(1) \\ &= n^{2/3} L Q_0(\pi_0, \theta_0) + O_p(1) \\ &= O_p(1)\end{aligned}$$

So

$$||\theta_0 - \hat{\theta}_n||_\infty = O_p\left(n^{-1/3}\right)$$

Step 2

Using an argument identical to the proof of Theorem 1 we can obtain:

$$\rho_c(\pi_0, \hat{\pi}_n^{pb})$$
$$= \sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\hat{\theta}_n}(t_i)} - E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right] \right| + \sum_{j=J_n+1}^{\infty} \eta_j \left| \langle \pi_0, \chi_j \rangle_W \right|$$

We can further break up the first sum into:

$$\sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\hat{\theta}_n}(t_i)} - E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right] \right|$$
$$\leq \sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\theta_0}(t_i)} - E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right] \right|$$
$$+ \sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\hat{\theta}_n}(t_i)} - \frac{1}{n\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\theta_0}(t_i)} \right|$$
$$= [A] + [B]$$

First we bound sum [A]. Using Lemma 5 we can replace the unconditional expectation with the ratio of conditional expectations:

$$\sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\theta_0}(t_i)} - E\left[\psi_j(T)\varphi_{\sigma_Y^2}(T)\right] \right|$$
$$= \sum_{j=0}^{J_n} \frac{1}{\lambda_j} \left| \frac{1}{n\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\theta_0}(t_i)} - \frac{E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)} \mid R\right]}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]} \right|$$

Assumption 2 guarantees that $w_{\theta_0}(t)$ is bounded away from zero. By the same

arguments as Theorem 1,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)} - E\left[\frac{1}{w_{\theta_0}(T)}\,|\,R\right] = O_p\left(n^{-1/2}\right)$$
$$\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\theta_0}(t_i)} - E\left[\frac{\psi_j(T)\,\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)}\,|\,R\right] = O_p\left(n^{-1/2}\right)$$

In the proof of Theorem 1 we showed that $\lambda_{J_n}\sum_{j=0}^{J_n}\frac{1}{\lambda_j} \to \frac{1}{1-\lambda_0}$. So we have:

$$[A] = O_p\left(\lambda_{J_n}^{-1}n^{-1/2}\right)$$

Bounding term [B] requires bounding the convergence of $\hat{\theta}_n$. We already proved in step 1 that $||\hat{\theta}_n - \theta_0||_\infty = O_p(n^{-1/3})$. Now by Assumption 3, $\sup_t\left|\frac{1}{w_{\hat{\theta}_n}(t)} - \frac{1}{w_{\theta_0}(t)}\right| = O_p\left(n^{-1/3}\right)$. This implies:

$$\frac{1}{n\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\hat{\theta}_n}(t_i)} - \frac{1}{n\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y^2}(t_i)}{w_{\theta_0}(t_i)} = O_p\left(n^{-1/3}\right)$$

Again using the fact that $\lambda_{J_n}\sum_{j=0}^{J_n}\frac{1}{\lambda_j} \to \frac{1}{1-\lambda_0}$,

$$[B] = O_p\left(\lambda_{J_n}^{-1}n^{-1/3}\right)$$

The regularization bias identical to the one in Theorem 1. So we have:

$$\rho_c(\pi_0, \hat{\pi}_n^{pb}) = O_p\left(\lambda_{J_n}^{-1}n^{-1/3} + \eta_{J_n}\right)$$

If the meta-analyst chooses J_n such that for some $c > 0$, $n^{1/3}\left(\frac{\sigma_Y^2}{\sigma_Y^2+1}\right)^{J_n/2} \to c$ then again by an identical argument to the one in the proof of Theorem 1, with $\frac{1}{2}$ exchanged

for $\frac{1}{3}$

$$\rho_c(\pi_0, \hat{\pi}_n^{pb}) = O_p\left(n^{-\frac{1}{2}\frac{\log\left(\frac{1+\sigma_Y^2}{1+\sigma_Y^2-c^{-2}}\right)}{\log\left(\frac{\sigma_Y^2+1}{\sigma_Y^2}\right)}}\right)$$

1.14 Additional Proofs

1.14.1 Proof of Theorem 12

We want to invoke Theorem 2.1 from Neumann (2013). To do this we need to show (i) weak dependence and the (ii) Lindebergh condition. Weak dependence is immediate because each observation is independent of all but C other observations and C is fixed. Next we check the Lindebergh condition. The following Lyapunov condition is sufficient for the Lindebergh condition:

$$\frac{E\left[\left(\sum_{j=0}^{J_n}\frac{1}{\lambda_j}Z_j(T,\theta_0)\right)^4\right]}{nE\left[\left(\sum_{j=0}^{J_n}\frac{1}{\lambda_j}Z_j(T,\theta_0)\right)^2\right]^2} \to 0$$

Next we show that the Lyapunov condition holds. Since $\lambda_j = \left(\frac{\sigma_Y^2}{1+\sigma_Y^2+c^{-2}}\right)^{j/2}$,we can bound the sum: $\lambda_{J_n}\sum_j^{J_n}\frac{1}{\lambda_j} = \sum_{j=0}^{J_n}\lambda_j < \frac{1}{1-\frac{\sigma_Y^2}{1+\sigma_Y^2+c^{-2}}}$. Combining this fact with (i) of Theorem 7, there is an $M > 0$ such that $\left|\sum_j \frac{1}{\lambda_j}Z_j(T,\theta_0)\right|/(Mn^{1/3}\lambda_{J_n}) < 1$. If we multiply the numerator and denominator by $M^{-4}\lambda_{J_n}^4 n^{-4/3}$, we obtain $\frac{E\left[\left(\sum_{j=0}^{J_n}\frac{\lambda_{J_n}}{\lambda_j}n^{-1/3}Z_j(T,\theta_0)/M\right)^4\right]}{nE\left[\left(\sum_{j=0}^{J_n}\frac{\lambda_{J_n}}{\lambda_j}n^{-1/3}Z_j(T,\theta_0)/M\right)^2\right]^2}$. Since $\left|\sum_{j=0}^{J_n}\frac{\lambda_{J_n}}{\lambda_j}n^{-1/3}Z_j(T,\theta_0)/M\right| < 1$, the fourth moment is smaller than the second moment. So, $\frac{E\left[\left(\sum_{j=0}^{J_n}\frac{1}{\lambda_j}Z_j(T,\theta_0)\right)^4\right]}{nE\left[\left(\sum_{j=0}^{J_n}\frac{1}{\lambda_j}Z_j(T,\theta_0)\right)^2\right]^2} \leq \frac{1}{n^{1/3}\lambda_{J_n}^2 E\left[\left(\sum_{j=0}^{J_n}\frac{1}{\lambda_j}Z_j(T,\theta_0)\right)^2\right]}$. By Assumption 5, the

sum of Z_j dominates the other terms and

$$\liminf n^{4/3}\lambda_{J_n}^2 n^{-1} E\left[\left(\sum_{j=0}^{J_n}\frac{1}{\lambda_j}Z_j(T,\theta_0)\right)^2\right] = \liminf n^{1/3}\lambda_{J_n}^2 E\left[\left(\sum_{j=0}^{J_n}\frac{1}{\lambda_j}Z_j(T,\theta_0)\right)^2\right] = \infty$$

So the Lyapunov condition holds.

1.14.2 Proof of Lemma 6

Using the statement from Appendix 1.12.2 for $\tilde{\Delta}_{c,n}$:

$$\begin{aligned}\tilde{\Delta}_{c,n} - \Delta_c = &\sum_{j=J_n+1}^{\infty}\eta_j\langle\psi_j,\pi_0\rangle_W a_j \\ &+\sum_{j=0}^{J_n}a_j\frac{E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T)\mid R\right]\left[G(\theta_0)'G(\theta_0)\right]^{-1}G(\theta_0)'E[\mathbf{v}(T)\mid R]}{\lambda_j E\left[\frac{1}{w_{\theta_0}(T)}\mid R\right]^2} \\ &+\frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)}\mid R\right]^2}B'_{\theta_0}\left[G(\theta_0)'G(\theta_0)\right]^{-1}G(\theta_0)'E[\mathbf{v}(T)\mid R]\end{aligned}$$

where a_j is shorthand for:

$$a_j \equiv \lambda_j\int_{-cv}^{cv}\phi_j(s)ds - \int_{-cv/c}^{cv/c}\phi_j(s)ds$$

First we bound the first sum. We showed in the Proof of Lemma 3 that the integrals $\int_{-cv/c}^{cv/c}\phi_j(s)ds$ and $\int_{-cv}^{cv}\phi_j(s)ds$ are bounded uniformly in j which means that $\sup_j|a_j|<\infty$. We also showed in the proof of Lemma 3 that $|\langle\psi_j,\pi_0\rangle_W|\leq 1$ for all j. We also showed in the proof of Theorem 1 that $\sum_{j=J_n+1}^{\infty}\frac{|\langle\psi_j,\pi_0\rangle_W|}{\lambda_j}\leq\eta_{J_n}$. By Holder's Inequality and the properties of geometric series, the regularization bias term is:

$$\sum_{j=J_n+1}^{\infty}\eta_j\langle\psi_j,\pi_0\rangle_W a_j = O\left(\sum_{j=J_n+1}^{\infty}\eta_j\right) = O(\eta_{J_n})$$

Next we bound the second and third sum. This involves bounding the $|\mathscr{X}| \times 1$ vector $E[\mathbf{v}(T) \mid R]$. It will be enough to bound its infinity norm since the number of element does not change. Recall from Appendix 1.12.3 that $\mathbf{v}_n(t)$ is a function that maps $t \in \mathbb{R}$ to a $|\mathscr{X}| \times 1$ vector with components:

$$\mathbf{v}_{n,x}(t) = \frac{1}{\varepsilon_n}\left(\frac{\mathbf{1}\{t \in (x, x+\varepsilon_n]\}}{w_{\theta_0}(x+\varepsilon_n)} - \frac{\mathbf{1}\{t \in (x-\varepsilon_n, x]\}}{w_{\theta_0}(x-\varepsilon_n)}\right)$$

For each x,

$$E[\mathbf{v}_{n,x}(T) \mid R] = \frac{1}{\varepsilon_n}\left(\frac{\Pr(T \in (x, x+\varepsilon_n] \mid R)}{w_{\theta_0}(x+\varepsilon_n)} - \frac{\Pr(T \in (x-\varepsilon_n, x] \mid R)}{w_{\theta_0}(x-\varepsilon_n)}\right)$$

Using the same argument as the proof in Lemma 5,

$$\begin{aligned}
\frac{1}{\varepsilon_n}\Pr(T \in (x, x+\varepsilon_n] \mid R) &= \mathrm{E}\left[\frac{\mathbf{1}\{T \in (x, x+\varepsilon_n]\}}{\varepsilon_n} w_{\theta_0}(T)\right] \frac{1}{E[w_{\theta_0}(T)]} \\
&= \mathrm{E}\left[\frac{\mathbf{1}\{T \in (x, x+\varepsilon_n]\}}{\varepsilon_n}\right] \frac{w_{\theta_0}(x+\varepsilon_n)}{E[w_{\theta_0}(T)]} \\
&\quad + \mathrm{E}\left[\frac{\mathbf{1}\{T \in (x, x+\varepsilon_n]\}}{\varepsilon_n}(w_{\theta_0}(T) - w_{\theta_0}(x+\varepsilon_n))\right] \frac{1}{E[w_{\theta_0}(T)]}
\end{aligned}$$

Since f_T is continuous,

$$\mathrm{E}\left[\frac{\mathbf{1}\{T \in (x, x+\varepsilon_n]\}}{\varepsilon_n}\right] \frac{1}{E[w_{\theta_0}(T)]} - \mathrm{E}\left[\frac{\mathbf{1}\{T \in [x-\varepsilon_n, x)\}}{\varepsilon_n}\right] \frac{1}{E[w_{\theta_0}(T)]} = O(\varepsilon_n)$$

The leaves the term: $\mathrm{E}\left[\frac{\mathbf{1}\{T \in (x, x+\varepsilon_n]\}}{\varepsilon_n}(w_{\theta_0}(T) - w_{\theta_0}(x+\varepsilon_n))\right] \frac{1}{E[w_{\theta_0}(T)]}$
Since $w_\theta(t)$ is by Assumption 3 uniformly continous except on a finite set,

$$\mathrm{E}\left[\frac{\mathbf{1}\{T \in (x, x+\varepsilon_n]\}}{\varepsilon_n}(w_{\theta_0}(T) - w_{\theta_0}(x+\varepsilon_n))\right] \frac{1}{E[w_{\theta_0}(T)]} = O(\varepsilon_n)$$

So we have: $E[\mathbf{v}_{n,x}(T) \mid R] = O(\varepsilon_n)$ and since the number of elements isn't changing

$||E[\mathbf{v}_n(T) \mid R]||_\infty = O(\varepsilon_n)$.

Now we bound the vectors that $E[\mathbf{v}_n(T) \mid R]$ is multiplied by. In the second sum, first recall that the integrals over ϕ are uniformly bounded over j. The matrix $[G(\theta_0)'G(\theta_0)]^{-1} G(\theta_0)'$ is not changing so we don't need to bound it. It remains to bound the sum of vector norms:

$$\sum_{j=0}^{J_n} \left|\left|E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right]\right|\right|_\infty$$

We can write this in terms of inner products:

$$\sum_{j=0}^{J_n} \left|\left|E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right]\right|\right|_\infty = \sum_{j=0}^{J_n} \left|\left|\left\langle \psi_j, f_{T|T} \frac{d}{d\theta_k} \frac{1}{w_\theta(t)} \right\rangle_Y \right|\right|_\infty$$

Assumption 3 says that $w_\theta(t)$ is everywhere differentiable in θ and that the derivatives are uniformly bounded. So $||f_{T|R} \frac{d}{d\theta_k} \frac{1}{w_\theta(t)}||_Y \leq ||f_{T|R}||_Y \leq 1$. So,

$$\sum_{j=0}^{J_n} \left|\left|E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right]\right|\right|_\infty \leq |\mathscr{X}|$$

In the third sum, $\frac{\beta_c}{E\left[\frac{1}{w_{\theta_0}(T)}|R\right]^2} B'_{\theta_0} [G(\theta_0)'G(\theta_0)]^{-1} G(\theta_0)'$ is a fixed vector.
Since all three terms are now bounded:

$$\tilde{\Delta}_{c,n} - \Delta_c = O(\eta_{J_n} + \varepsilon_n)$$

1.14.3 Proof of Lemma 7

Since it minimizes the sample objective at zero, $\hat{\pi}_n$ must satisfy:

$$\eta_j \langle \chi_j, \hat{\pi}_n \rangle_W = \frac{1}{\lambda_j} \frac{\frac{1}{n}\sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)}}{\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}}$$

Let the deterministic sequence $r_{1,n}$ be:

$$r_{1,n} \equiv \sum_{j=J_n+1}^{\infty} \eta_j \langle \psi_j, \pi_0 \rangle_W \left(\lambda_j \int_{-cv}^{cv} \phi_j(s)ds - \int_{-cv/c}^{cv/c} \phi_j(s)ds \right)$$

Using Taylor's Theorem for some $w_{\bar{\theta}_n}$ such that $E\left[\frac{1}{w_{\bar{\theta}}(T)}\right]$ is between $E\left[\frac{1}{w_{\hat{\theta}_n}(T)}\right]$ and $E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]$:

$$\begin{aligned}
&\frac{1}{\lambda_j} \frac{\frac{1}{n}\sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)}}{\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)}} \\
&= \frac{1}{\lambda_j} \frac{\frac{1}{n}\sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)}}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]} \\
&- \frac{1}{E\left[\frac{1}{w_{\bar{\theta}}(T)}\right]^2} \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)} - E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right] \right) \frac{1}{\lambda_j} \frac{1}{n} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)} \\
&= \frac{1}{\lambda_j} \frac{\frac{1}{n}\sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)}}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]} \\
&- \frac{1}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)} - E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right] \right) \frac{1}{\lambda_j} \frac{1}{n} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)} \\
&- \frac{1}{E\left[\frac{1}{w_{\bar{\theta}}(T)} \mid R\right]^2} \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\hat{\theta}_n}(t_i)} - \frac{1}{n}\sum_{i=1}^n \frac{1}{w_{\theta_0}(t_i)} \right) \frac{1}{\lambda_j} \frac{1}{n} \sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)}
\end{aligned}$$

Since $\left|\frac{1}{w_{\hat{\theta}_n}(t)} - \frac{1}{w_{\theta_0}(t)}\right| = O_p\left(n^{-1/3}\right)$ and $w_\theta(t)$ is uniformly bounded away from zero and uniformly bounded above by Assumption 2, $\frac{1}{E\left[\frac{1}{w_{\bar{\theta}}(T)}\right]^2} - \frac{1}{E\left[\frac{1}{w_{\theta_0}(T)}|R\right]^2} = O\left(n^{-1/3}\right)$. We have already shown that $Var(\psi_j(T)\varphi_{\sigma_Y}(T)) \leq 1$ so $\frac{1}{n}\sum_{i=1}^n \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)} - E\left[\frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\theta_0}(t_i)}\right] = O_p(n^{-1/3})$. So we have:

$$\begin{aligned}
\eta_j\langle\chi_j,\hat{\pi}_n\rangle_W - \eta_j\langle\chi_j,\pi_0\rangle_W =& \frac{1}{\lambda_j}\frac{\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)} - E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)}\right]}{E\left[\frac{1}{w_{\theta_0}(T)}\mid R\right]} \\
&- \frac{E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)}\right]}{E\left[\frac{1}{w_{\theta_0}(T)}\mid R\right]^2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\theta_0}(t_i)} - E\left[\frac{1}{w_{\theta_0}(T)}\mid R\right]\right)\frac{1}{\lambda_j} \\
&- \frac{E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)}\right]}{E\left[\frac{1}{w_{\theta_0}(T)}\mid R\right]^2}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)} - \frac{1}{w_{\theta_0}(t_i)}\right)\frac{1}{\lambda_j} \\
&+ O_p\left(\frac{1}{\lambda_j n^{1/3}}\right) \\
=&[D]+[E]+[F]+O_p\left(\frac{1}{\lambda_j n^{2/3}}\right)
\end{aligned}$$

Term [E] can already be written as a sample mean minus its expectation. For the other two terms, we need to establish the following facts. We will use the Taylor Theorem. To do so, first define the following two $K\times 1$ gradient vectors:

$$\begin{aligned}
\hat{B}_\theta(t_i) &\equiv \nabla_\theta \frac{1}{w_\theta(t_i)} \\
B_\theta &\equiv E\left[\hat{B}_{\theta_0}(T)\mid R\right]
\end{aligned}$$

Since the gradients and θ_0 are $K\times 1$, $B'_{\theta_0}\theta_0$ is a scalar. With the gradients defined, we can use the Taylor Theorem:

$$\frac{1}{w_{\hat{\theta}_n}(t_i)} - \frac{1}{w_{\theta_0}(t_i)} = B'_{\theta_0}(t_i)\left(\hat{\theta}_n-\theta_0\right) + O_p\left(\left(\hat{\theta}_n-\theta_0\right)^2\right)$$

Since we assumed that each element of B_θ is uniformly bounded:

$\left|\left|\frac{1}{n}\sum_{i=1}^{n}\hat{B}_{\theta}(t_i) - E\left[\hat{B}_{\theta_0}(T) \mid R\right]\right|\right|_{\infty} = O_p\left(n^{-1/2}\right)$. This gives us:

$$[F] = \frac{1}{\lambda_j}\frac{E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)} \mid R\right]}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2}B'_{\theta_0}\left(\hat{\theta}_n - \theta_0\right) + O_p\left(\frac{1}{n^{2/3}\lambda_j}\right)$$

To linearize term [D]. First consider its numerator.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)} - E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)} \mid R\right] &= \frac{1}{n}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\theta_0}(t_i)} - E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)} \mid R\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\hat{\theta}_n}(t_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\theta_0}(t_i)} \\ &= [D1] + [D2]\end{aligned}$$

Term [D1] is a sample mean minus its expectation. Term [D2] can be written as:

$$\begin{aligned}[D2] &= \frac{1}{n}\sum_{i=1}^{n}\psi_j(t_i)\varphi_{\sigma_Y}(t_i)\left(\frac{1}{w_{\hat{\theta}_n}(t_i)} - \frac{1}{w_{\theta_0}(t_i)}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\psi_j(t_i)\varphi_{\sigma_Y}(t_i)\hat{B}'_{\theta_0}(t_i)(\hat{\theta}_n - \theta_0) + O_p\left((\hat{\theta}_n - \theta_0)^2\right) \\ &= E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right](\hat{\theta}_n - \theta_0) + O_p\left(n^{-2/3}\right)\end{aligned}$$

To keep things concise we will need to define the following shorthand. We showed in the proof of Lemma 3 that the integrals are uniformly bounded over all j. Since the λ_j are decreasing this means that $\sup_j |a_j| < \infty$.

$$a_j \equiv \lambda_j \int_{-cv}^{cv}\phi_j(s)ds - \int_{-cv/c}^{cv/c}\phi_j(s)ds$$

No we can write the entire esimation error as:

$$\begin{aligned}\hat{\Delta}_{c,n} - \Delta_c + r_{1,n} = & \frac{1}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\theta_0}(t_i)} - \sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)} \mid R \right] \right) \\ & + \frac{\sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R \right]}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} (\hat{\theta}_n - \theta_0) \\ & + \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{w_{\theta_0}(t_i)} - E\left[\frac{1}{w_{\theta_0}(T)} \mid R \right] \right) \\ & + \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0} \left(\hat{\theta}_n - \theta_0 \right) \\ & + O_p\left(\frac{1}{\lambda_{J_n} n^{2/3}} \right)\end{aligned}$$

To proceed we must now linearize $\hat{\theta}_n - \theta_0$. Using the result from section 1.14.3 below, where the $|\mathscr{X}| \times K$ matrix G and $K \times 1$ vector $\hat{g}_n$ are defined there:

$$\hat{\theta}_n - \theta_0 = \left[G(\theta_0)' G_n(\theta_0) \right]^{-1} G(\theta_0)' n^{1/3} \hat{g}_n(\theta_0) + O_p\left(n^{-2/3} \right)$$

Substituting this into the first term and using the facts that $\sup_j |a_j| < \infty$ and $\lambda_J \sum_{j=0}^{J} \lambda_j^{-1} = O(1)$:

$$\begin{aligned}& \frac{\sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R \right]}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} (\hat{\theta}_n - \theta_0) \\ & = \frac{\sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R \right] \left[G(\theta_0)' G_n(\theta_0) \right]^{-1} G(\theta_0)' \hat{g}_n(\theta_0)}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} + O_p\left(\frac{1}{\lambda_{J_n} n^{2/3}} \right)\end{aligned}$$

Linearizing the second term:

$$\frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0}\left(\hat{\theta}_n - \theta_0\right) = \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0}\left[G(\theta_0)' G_n(\theta_0)\right]^{-1} G(\theta_0)' \hat{g}_n(\theta_0) + O_p\left(n^{-2/3}\right)$$

Putting all of these pieces together:

$$\begin{aligned}
&\hat{\Delta}_{c,n} - \Delta_c + r_{1,n} \\
&= \frac{1}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} \frac{\psi_j(t_i)\varphi_{\sigma_Y}(t_i)}{w_{\theta_0}(t_i)} - \sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\frac{\psi_j(T)\varphi_{\sigma_Y}(T)}{w_{\theta_0}(T)} \mid R \right] \right) \\
&+ \frac{\sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right] \left[G(\theta_0)' G_n(\theta_0)\right]^{-1} G(\theta_0)' \hat{g}_n(\theta_0)}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \\
&+ \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{1}{w_{\theta_0}(t_i)} - E\left[\frac{1}{w_{\theta_0}(T)} \mid R \right] \right) \\
&+ \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0}\left[G(\theta_0)' G_n(\theta_0)\right]^{-1} G(\theta_0)' \hat{g}_n(\theta_0) \\
&+ O_p\left(\frac{1}{\lambda_{J_n} n^{2/3}} \right)
\end{aligned}$$

Now we write down the function Z_{n,J_n} explicitly:

$$\begin{aligned}
Z_{n,J_n}(t) &= \frac{1}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]} \sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} \frac{\psi_j(t)\varphi_{\sigma_Y}(t)}{w_{\theta_0}(t)} \\
&+ \frac{\sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right] \left[G(\theta_0)'G_n(\theta_0)\right]^{-1} G(\theta_0)'\mathbf{v}(t)}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \\
&+ \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} \frac{1}{w_{\theta_0}(t)} \\
&+ \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0} \left[G(\theta_0)'G_n(\theta_0)\right]^{-1} G(\theta_0)'\mathbf{v}(t) \\
&= [G] + [H] + [I] + [J]
\end{aligned}$$

Where $\mathbf{v}_n(t)$ is an $|\mathscr{X}| \times 1$ vector with components:

$$\mathbf{v}_{n,x}(t) = \frac{1}{\varepsilon_n} \left(\frac{\mathbf{1}\{t \in (x, x+\varepsilon_n]\}}{w_{\theta_0}(x+\varepsilon_n)} - \frac{\mathbf{1}\{t \in (x-\varepsilon_n, x]\}}{w_{\theta_0}(x-\varepsilon_n)} \right)$$

We will establish two facts:

- $\sup_{t,n} n^{-1/3}\lambda_{J_n}|Z_{n,J_n}(t)| < \infty$.
- $\sup_n \lambda_{J_n}^{-2} n^{-1/3} Var(Z_{n,J_n}(T)) < \infty$

First we show that $\sup_{t,n} n^{-1/3}\lambda_{J_n}|Z_{n,J_n}(t)| < \infty$. First, consider [G]. The functions $\psi_j(t)\varphi_{\sigma_Y}(t)$ are called the "Hermite Functions." They are uniformly bounded by Cramér's Inequality which can be found in Indritz (1961) and Bateman (1953):

$$\sup_{j,t} |\psi_j(t)\varphi_{\sigma_Y}(t)| < \pi^{1/2}$$

We have already shown that $\lambda_{J_n}\sum_{j=0}^{J_n}\frac{1}{\lambda_j}=O(1)$. So we can conclude:

$$[G]\lesssim\frac{1}{\lambda_{J_n}}$$

For the same reasons:

$$\sum_{j=0}^{J_n}\frac{a_j}{\lambda_j}E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T)\mid R\right]\left[G(\theta_0)'G_n(\theta_0)\right]^{-1}G(\theta_0)'\lesssim\frac{1}{\lambda_{J_n}}$$

By Assumption 2,

$$\mathbf{v}_{n,x}(t)\lesssim\frac{1}{\varepsilon_n}\lesssim n^{1/3}$$

So we conclude that $\sup_t|[H]|\lesssim\frac{n^{1/3}}{\lambda_{J_n}}$. Assumption 2 guarantees that $\sup_t|[I]|\lesssim 1$. Finally, using the same arguments as the for the other parts: $\sup_t|[J]|\lesssim n^{1/3}$. Putting all three parts together,

$$\sup_{t\in\mathbb{R}}\max_{j\in 1,\cdots J_n}|Z_{j,J_n}(t)|\lesssim\frac{n^{1/3}}{\lambda_{J_n}}$$

Now to show $\sup_n\lambda_{J_n}^{-2}n^{-1/3}Var(Z_{n,J_n}(T))<\infty$. The variance of the sum of four terms is no greater than sixteen times the largest of their variances. So we need only show that [G], [H], [I], and [J] all have variances of order $\lambda_{J_n}^{-2}n^{-1/3}$. For [G] and [I], the upper bounds already established suffice since the variance cannot be greater than the square of the maximum. For [H] and [J], it is sufficient to establish that $\max_x Var(\mathbf{v}_{n,x}(T))\lesssim n^{1/3}$. To do this, first use Assumption 2 to bound $1/w_{\theta_0}$. Then notice that $\mathbf{v}_{n,x}(T)$ is $n^{1/3}$ times a bounded number times the difference of two Bernoulli random variables. The variance of a Bernoulli is $p(1-p)$. Since T must have a continuous distribution $P(T\in(x,x+\varepsilon_n])\lesssim\varepsilon_n\lesssim\frac{1}{n^{1/3}}$. So $Var(\mathbf{v}_{n,x}(T))\lesssim\frac{n^{2/3}}{n^{1/3}}=n^{1/3}$. Since there are only finitely many x, $\max_x Var(\mathbf{v}_{n,x}(T))\lesssim n^{1/3}$ and we have established that [H],[J] have variance of order $\lambda_{J_n}^{-2}n^{-1/3}$, which is enough to guarantee that $\sup_n\lambda_{J_n}^{-2}n^{-1/3}Var(Z_{n,J_n}(T))<\infty$.

Notice that $\frac{1}{n}\sum_{i=1}^{n}Z_{n,J_n}(t_i)-E[Z_{n,J_n}(T)]$ is not asymptotically equivalent to $\hat{\beta}_{c,n}-$

$\tilde{\beta}_{c,n}$ because $E[\mathbf{v}_{n,x}(t)] \neq 0$. We need to subtract this bias term off on the left-hand side.

$$E\left[\mathbf{v}_{n,x}(t)\right] = \frac{1}{\varepsilon_n}\left(\frac{P\left(t \in (x, x+\varepsilon_n]\right)}{w_{\theta_0}(x+\varepsilon_n)} - \frac{P\left(t \in (x-\varepsilon_n, x]\right)}{w_{\theta_0}(x-\varepsilon_n)}\right)$$

Now define the bias terms as:

$$r_{2,n} \equiv \frac{\sum_{j=0}^{J_n} \frac{a_j}{\lambda_j} E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right]\left[G(\theta_0)'G_n(\theta_0)\right]^{-1} G(\theta_0)'E[\mathbf{v}(T) \mid R]}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2}$$
$$+ \frac{\Delta_c}{E\left[\frac{1}{w_{\theta_0}(T)} \mid R\right]^2} B'_{\theta_0}\left[G(\theta_0)'G_n(\theta_0)\right]^{-1} G(\theta_0)'E[\mathbf{v}(T) \mid R]$$

This expectation is $E[||\mathbf{v}(T)||_\infty] = O(\varepsilon_n) = O\left(n^{-1/3}\right)$. So we can conclude:

$$\hat{\Delta}_{c,n} - \Delta_0 - r_{1,n} - r_{2,n} = \frac{1}{n}\sum_{i=1}^{n} Z_{n,J_n}(t_i) - E\left[Z_{n,J_n}(T)\right] + O_p\left(n^{-2/3}\lambda_{J_n}\right)$$

Linearization of $\hat{\theta}_n - \tilde{\theta}_n$

Next we study the convergence of $\hat{\theta}_n$. Since the first term of the sample objective is zero at its minimizer with probability approaching 1, $\hat{\theta}_n$ satisfies

$$\hat{\theta}_n = \arg\min_{\theta} \sum_{x \in \mathscr{X}} \left(\frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^{n} \mathbf{1}\left\{t_i \in (x, x+\varepsilon_n]\right\}}{w_\theta(x+\varepsilon_n)} - \frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^{n} \mathbf{1}\left\{t_i \in (x-\varepsilon_n, x]\right\}}{w_\theta(x-\varepsilon_n)}\right)^2$$

This form coincides with a classical minimum distance estimator and we can show its asymptotic normality in the standard way from Newey and McFadden (1994). Let $\hat{g}_n(\theta)$ be a $|\mathscr{X}| \times 1$ vector with components equal to the summands:

$\left(\frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i\in(x,x+\varepsilon_n]\}}{w_\theta(x+\varepsilon_n)} - \frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i\in(x-\varepsilon_n,x]\}}{w_\theta(x-\varepsilon_n)}\right)$. So,

$$\hat{\theta}_n = \arg\min_\theta \hat{g}_n(\theta)'\hat{g}_n(\theta)$$

Since $w_\theta(t)$ is assumed to be everywhere continuously differentiable in θ, $\hat{g}_n(\theta)$ is also everywhere continuously differentiable in θ. Let $\hat{G}(\theta)$ be the $|\mathscr{X}|\times K$ matrix of first derivatives of $\hat{g}_n(\theta)$ with respect to the K components of θ.

$$\begin{aligned}&\hat{G}(\theta)_{x,k}\\ &= -\frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i\in(x,x+\varepsilon_n]\}}{w_\theta(x+\varepsilon_n)^2}\frac{d}{d\theta}w_\theta(x+\varepsilon_n) + \frac{\frac{1}{n\varepsilon_n}\sum_{i=1}^n \mathbf{1}\{t_i\in(x-\varepsilon_n,x]\}}{w_\theta(x-\varepsilon_n)^2}\frac{d}{d\theta}w_\theta(x-\varepsilon_n)\end{aligned}$$

The population gradient $G(\theta)$ is:

$$G(\theta)_{x,k} = -\lim_{\varepsilon\to 0}\frac{f_{T|R}(x+\varepsilon)}{w_\theta(x+\varepsilon_n)^2}\frac{d}{d\theta}w_\theta(x+\varepsilon) + \frac{f_{T|R}(x-\varepsilon)}{w_\theta(x-\varepsilon)^2}\frac{d}{d\theta}w_\theta(x-\varepsilon)$$

The sample gradients converge uniformly to the population gradient:

$$\max_{x\in\mathscr{X},k\in\{1,\cdots K\}}\sup_\theta n^{1/3}\left|\hat{G}(\theta)_{x,k} - G(\theta)_{x,k}\right| = O_p(1)$$

So at the solution the FOC is satisfied:

$$\mathbf{0} = \hat{G}(\hat{\theta}_n)'\hat{g}_n(\hat{\theta}_n)$$

Taking the Taylor expansion about $\tilde{\theta}$:

$$\hat{g}_n(\hat{\theta}_n) - \hat{g}_n(\theta_0) = \hat{G}_n(\bar{\theta}_n)(\hat{\theta}_n - \tilde{\theta}_n)$$

Substituting this into the FOC:

$$\mathbf{0} = \hat{G}(\hat{\theta}_n)'\hat{g}_n(\theta_0) + \hat{G}(\hat{\theta}_n)'\hat{G}_n(\bar{\theta}_n)(\hat{\theta}_n - \tilde{\theta}_n)$$

Solving and using the nonsingularity of G,

$$(\hat{\theta}_n - \tilde{\theta}_n) = \left[\hat{G}(\hat{\theta}_n)'\hat{G}_n(\bar{\theta}_n)\right]^{-1}\hat{G}(\hat{\theta}_n)'\hat{g}_n(\theta_0)$$

Since $(\tilde{\theta}_n - \theta_0) \to_p 0$ and $\overline{\theta}$ is a mean value and the derivatives of w_θ in θ are continuous and the fact that matrix inverses are continuous in the 2-norm at invertible matrices:

$$\left|\left|\left[G(\theta_0)'G_n(\theta_0)\right]^{-1}G(\theta_0)' - \left[\hat{G}(\hat{\theta}_n)'\hat{G}_n(\bar{\theta}_n)\right]^{-1}\hat{G}(\hat{\theta}_n)'\right|\right|_2 \to_p 0$$

Since the dimensions of G are fixed and each dimension is converging with $n^{-1/3}$, $n^{1/3}||\hat{G}(\hat{\theta}_n) - G(\theta_0)||_F = O_p(1)$. Since the Frobenius norm upper bounds the 2-norm, $n^{1/3}||\hat{G}(\hat{\theta}_n) - G(\theta_0)||_2 = O_p(1)$. Since matrix multiplication and inversion are Lipschitz continuous with respect to the 2-norm:

$$n^{1/3}\left|\left|\left[G(\theta_0)'G_n(\theta_0)\right]^{-1}G(\theta_0)' - \left[\hat{G}(\hat{\theta}_n)'\hat{G}_n(\bar{\theta}_n)\right]^{-1}\hat{G}(\hat{\theta}_n)'\right|\right|_2 = O_p(1)$$

Since $n^{1/3}\hat{g}_n(\theta_0) = O_p(1)$,

$$n^{1/3}(\hat{\theta}_n - \tilde{\theta}_n) = \left[G(\theta_0)'G_n(\theta_0)\right]^{-1}G(\theta_0)'n^{1/3}\hat{g}_n(\theta_0) + O_p\left(n^{-1/3}\right)$$

But $\hat{g}_n(\theta_0)$ is a vector of sample averages.

1.14.4 Proof of Theorem 4

Define:

$$\begin{aligned}\hat{Z}_{n,J_n}(t) = &\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}}\sum_{j=0}^{J_n}\frac{a_j}{\lambda_j}\frac{\psi_j(t)\varphi_{\sigma_Y}(t)}{w_{\hat{\theta}_n}(t)} \\ &+\frac{\sum_{j=0}^{J_n}\frac{a_j}{\lambda_j}E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\hat{\theta}_n}(T)\mid R\right]\left[G(\hat{\theta}_n)'G_n(\hat{\theta}_n)\right]^{-1}G(\hat{\theta}_n)'\mathbf{v}_{\hat{\theta}_n}(t)}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}\right)^2} \\ &+\frac{\hat{\Delta}_{c,n}}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}\right)^2}\frac{1}{w_{\hat{\theta}_n}(t)} \\ &+\frac{\hat{\Delta}_{c,n}}{\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}\right)^2}B'_{\hat{\theta}_n}\left[G(\hat{\theta}_n)'G_n(\hat{\theta}_n)\right]^{-1}G(\hat{\theta}_n)'\mathbf{v}_{\hat{\theta}_n}(t)\end{aligned}$$

The proof proceeds in two steps. First we show that $\sup_{t\in\mathbb{R}} n^{1/3}\lambda_{J_n}|\hat{Z}_{n,J_n}(t) - Z_{n,J_n}(t)| = O_p(1)$. Then we show that the variance estimator converges faster than the variance decays.

Step 1

First we want to show that: $\sup_{t\in\mathbb{R}} n^{1/3}\lambda_{J_n}|\hat{Z}_{n,J_n}(t) - Z_{n,J_n}(t)| = O_p(1)$. We will show that each term above converges at this rate one by one.

First, we have already shown in Theorem 1 that $\hat{\Delta}_{c,n} - \Delta_c = O_p\left(\lambda_{J_n}n^{-1/3}\right)$. We showed in step 1 of the proof of Theorem 2 that $\sup_t n^{1/3}\left|\frac{1}{w_{\hat{\theta}_n}(t)} - \frac{1}{w_{\theta_0}(t)}\right| = O_p(1)$. This implies that $\sup_t n^{1/3}||\mathbf{v}_{\hat{\theta}_n}(t) - \mathbf{v}_{\hat{\theta}_0}(t)||_\infty = O_p(1)$. By the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)} - E\left[\frac{1}{w_{\theta_0}(T)}\right] = O_p\left(n^{-1/2}\right)$.

By Assumption 2, $\frac{1}{w_\theta(t)} > \overline{\theta}$ for all t,θ, so $\left|\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}} - E\left[\frac{1}{w_{\theta_0}(T)}\right]\right| = O_p(n^{-1/3})$. Since all quantities are uniformly bounded by $\overline{\theta}$, $\left|\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_{\hat{\theta}_n}(t_i)}} - E\left[\frac{1}{w_{\theta_0}(T)}\right]\right|^2 = O_p(n^{-1/3})$. Next, we showed in the proof of Lemma 7 that $n^{1/3}\left|\left|\left[G(\hat{\theta}_n)'G_n(\hat{\theta}_n)\right]^{-1}G(\hat{\theta}_n)'\right|\right|_2 = O_p(1)$. Finally, since we assumed that $\nabla_\theta\frac{1}{w_\theta(t)}$ is Lipschitz-continuous in θ, $\sup_t n^{1/3}\left|\left|\hat{B}_{\hat{\theta}_n}(t) - \right|\right|_1 <$

∞. Combining this with the fact that $\psi_j(t)\varphi_{\sigma_Y}(t) < \pi^{1/4}$ from Indritz (1961),

$$\left|\left|E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\hat{\theta}_n}(T) \mid R\right] - E\left[\psi_j(T)\varphi_{\sigma_Y}(T)\hat{B}'_{\theta_0}(T) \mid R\right]\right|\right|_1 = O\left(n^{-1/3}\right)$$

Combining all of these facts, we conclude that $\sup_{t\in\mathbb{R}} n^{1/3}\lambda_{J_n}|\hat{Z}_{n,J_n}(t) - Z_{n,J_n}(t)| = O_p(1)$.

Step 2

Now we show the main result. $n^{5/6}\lambda_{J_n}^2$ times the difference between variance estimator and the variance estimator if we knew Z_{n,J_n} is bounded by:

$$\begin{aligned}
&n^{5/6}\lambda_{J_n}^2\left|\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}Z_{n,J_n}(t_i)Z_{n,J_n}(t_k) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}\hat{Z}_{n,J_n}(t_i)\hat{Z}_{n,J_n}(t_k)\right| \\
&= O_p\left(n^{5/6}\lambda_{J_n}^2\frac{1}{n}\right) \\
&= O_p\left(n^{1/6}\lambda_{J_n}^2\right) \\
&= o_p(1)
\end{aligned}$$

But by Assumption 5, $\liminf n^{5/6}\lambda_{J_n}^2 Var(\hat{\Delta}_{n,c}) = \infty$, so this difference must be dominated by the variance itself.

Now consider the estimation error.

Since $Var\left(Z_{n,J_n}(t_i)\right) = O(\lambda_{J_n}^{-2}n^{1/3})$, then $Var\left(Z_{n,J_n}(t_i)^2\right) = O(\lambda^{-4}n^{4/3})$.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}Z_{n,J_n}(t_i)Z_{n,J_n}(t_k) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}E[Z_{n,J_n}(t_i)Z_{n,J_n}(t_k)] &= O_p\left(\frac{n^{2/3}\lambda_{J_n}^{-2}}{n^{1/2}}\right) \\
&= O_p\left(n^{1/6}\lambda_{J_n}^{-2}\right)
\end{aligned}$$

So dividing by a further n yields:

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}Z_{n,J_n}(t_i)Z_{n,J_n}(t_k) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}E[Z_{n,J_n}(t_i)Z_{n,J_n}(t_k)] = O_p\left(n^{-5/6}\lambda_{J_n}^{-2}\right)$$

Multiplying by $n^{5/6}\lambda_{J_n}^2$:

$$n^{5/6}\lambda_{J_n}^2\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}Z_{n,J_n}(t_i)Z_{n,J_n}(t_k)-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}E[Z_{n,J_n}(t_i)Z_{n,J_n}(t_k)]\right)=O_p(1)$$

Again by Assumption 5, $\liminf n^{5/6}\lambda_{J_n}^2 Var(\hat{\Delta}_{n,c})=\infty$, so this difference is also dominated by the variance itself. So we can conclude,

$$\frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\Lambda_{ik}\hat{Z}_{n,J_n}(t_i)\hat{Z}_{nJ_n}(t_k)}{Var(\hat{\Delta}_{n,c})}\rightarrow_p 1$$

Chapter 1, in part is currently being prepared for submission for publication of the material. The dissertation author was the sole author of chapter 1.

Chapter 2

Linear Estimation of Global Average Treatment Effects

with Paul Niehaus

2.1 Introduction

Experimental program evaluations typically treat a subset of the relevant units in order to determine whether or not more of them should then be treated. Miguel and Kremer (2004)'s results on the impacts of deworming medication in a sample of 2,300 schoolchildren in Kenya, for example, helped shaped subsequent debate (the so-called "worm wars") over whether or not to administer similar medication to *all* schoolchildren in Kenya, and beyond. In such settings a parameter of interest is the "global average treatment effect" (GATE), i.e. the (average) causal effect of treating all units within some given class under consideration vs treating none of them.

Learning the GATE is challenging because it compares two counterfactual states—those in which everyone and no one is treated—neither of which is directly observed in any experiment. Some sort of extrapolation is necessary. The key to this turns out to be how to handle interference or "spillovers" between units. In the past the predominant approach has been to assume that any spillovers are "local" in some sense, so that

treating one unit affects only a small number of neighboring units (Manski, 2013).[1] For example, one might assume that spillovers are contained within villages, in which case comparing fully treated to fully untreated villages yields consistent estimates of the GATE. But such assumptions are often in uncomfortable tension with economic logic. Equilibrium models in economics typically have the property that everyone's behavior affects everyone else to some degree—through prices, unpriced externalities, strategic interactions, and so on. Recent field experiments conducted at relatively large scales have illustrated precisely these kinds of far-reaching general-equilibrium effects (Muralidharan and Niehaus, 2017). Reforming a public workfare program in India, for example, affected market wages and employment, land rents, and firm entry, among other outcomes (Muralidharan et al., 2021). Causal channels like these seem hard to reconcile with strictly local spillovers.

The result is that practice in this area has moved ahead of its theoretical underpinnings. Recent analyses of large-scale field experiments have specified estimators based on OLS regressions of a unit's outcomes on both its own treatment status and the mean treatment status of other units within some specified distance (e.g., Miguel and Kremer, 2004; Muralidharan et al., 2021; Egger et al., 2022). There is of course no more reason to believe that economic spillovers stop after some fixed number of kilometers than to believe they stop at administrative boundaries. These estimators do raise the hope that one might achieve asymptotic consistency by growing the relevant distances with the sample. But it is currently unclear whether and under what conditions this is the case, including in particular what is required of the experimental design. Regression-based estimators also embed an assumption that potential outcomes are linear in treatment, and it is unclear whether this is necessary for consistency or beneficial for performance. Overall, it is unclear how to choose designs and estimators likely to perform well.

[1]The "Stable Unit Treatment Value Assumption" (SUTVA) assumption is a very strong version of this, implying that *no* other units are affected.

This paper examines what can be achieved under the assumption that the effect on one unit of treating others decays with the "distance" between them. This idea seems intuitive enough; it is part of the justification for the estimators mentioned above, and also appears to be consistent with modern spatial general equilibrium modelling techniques in economics. Endogenous economic interactions decay geometrically with distance in the gravity (Allen et al., 2020) and market access (Donaldson and Hornbeck, 2016) traditions, for example. Defining a concept of "distance" appropriate for causal analysis, however, turns out to be somewhat subtle. The "distance" between two units that governs their economic interactions need not be a proper metric even if they themselves interact in Euclidean space—a point we illustrate with a simple economic example. Given this, we work with a generalized notion of distance that can be asymmetric and must satisfy only a loosened version of the triangle inequality. Our results are thus available both to practitioners comfortable making reduced form assumptions about the spatial decay of spillovers, and also to those working from a foundation of equilibrium theory.

For concreteness we work with a bound on spillovers that decays with at least a geometric rate $\eta > 0$. We pair this assumption with mild requirements on population dynamics that ensure the population does not become arbitrarily dense or dispersed as it grows. We also allow for arbitrary cluster-randomized designs in order to study performance with respect to the choice of design as well as estimator.

Our first set of results characterizes the optimal rate of convergence across all estimators that are linear in the outcomes with bounded weights and across all cluster-randomized designs, and shows that it is achievable. No such estimator can be guaranteed to converge to the GATE at a geometric rate faster than $n^{-\frac{1}{2+1/\eta}}$. The fact that this quantity increases in η illustrates the intuitive idea that the GATE is easier to estimate when the population is economically less interconnected (so that η is large). At the same time, the fact that estimators cannot converge at the $n^{-\frac{1}{2}}$ parametric rate quantifies one sense in which the problem is relatively difficult. The optimal rate can be achieved using inverse

probability weighting (IPW)-based estimators (a standard Horvitz-Thompson estimator or a refinement in the spirit of Hajek (1971)) paired with an experimental design from a "Scaling Clusters" class which we define. This class is broad enough to encompass many designs used in practice, and the conditions that define it are straight-forward to check. The essential idea is to let clusters grow with the population in such a way that the probability of any one unit having *all* of its neighbors (within some growing radius) treated or untreated is bounded below. These initial results build on and extend an important recent contribution by Leung (2022b), who demonstrates that when using a Horvitz-Thompson estimator in $\mathbb{R}^2$ space the optimal rate is achievable using clusters that take the form of growing squares if $\eta > 1$. We extend this result to non-Euclidean spaces required for economic applications, optimize over estimators as well as designs, and allow $\eta \in (0, \infty)$.

While the latter feasibility result is positive, it does not apply directly to existing applied work. None of the applied papers referenced above use IPW estimators; all use estimators based on OLS regressions that include the share of treated neighbors as a regressor. This may reflect the fact that IPW estimators use only a small part of the data and are thus thought to perform poorly in small samples, especially in experiments with small clusters. Regardless, it raises the questions whether a linear regression can yield consistent estimates of the GATE, how exactly it must be designed to do so, and whether this does in fact yield performance benefits.

To address these questions we introduce a second core assumption, implicit in regression-based estimators, of linear causal effects: potential outcomes are linear in the treatment status of all units. The effect on i of treating j is thus independent of the treatment status of k. We also show that it is sufficient for the existence of a consistent regression-based estimator. Here, however, the details matter. Estimators like those in the literature based on a regression of outcomes on the share of treated neighbors, for example, concentrate around a weighted average of the treatment effects of interest, not

around the (unweighted) GATE. The weights are positive, and in this sense the problem is less dire than that recently identified in two-way fixed effect designs (de Chaisemartin and D'Haultfœuille, 2020; Goodman-Bacon, 2021). However, since the weights depend on both the treated unit and the affected unit, they have no welfarist interpretation. This problem is relatively straight-forward to correct: a regression of outcomes on the number of treated clusters in a neighborhood (divided by the mean number of clusters across all neighborhoods) yields a consistent estimator of the unweighted GATE.[2]

As for performance, the assumption of linear causal effects has no impact on the best guaranteeable geometric rate of convergence as long as the experimenter can control cluster size; this remains $n^{-\frac{1}{2+1/\eta}}$, and can be achieved using our recommended regression-based estimator paired with a Scaling Clusters design. Where linear potential outcomes shine is in settings where the researcher is constrained to run an experiment with clusters that are (asymptotically) small; here OLS can be made to converge faster than Horvitz-Thompson or Hajak estimators. In this sense one can view the linear potential outcomes assumption as a substitute for the freedom to choose a design with large clusters. Regression-based approaches are also amenable to the introduction of additional prior information the researcher may have about patterns of economic interaction—who usually shops in which markets, for example, or which firms buy and sell from each other. As an extension we illustrate how such information can be incorporated using a two-stage-least-square approach and (potentially) improve performance.

Our third set of results return to the world of nonlinear potential outcomes and provide guidance regarding which approaches are likely to perform well in a given finite-sample study. Generally speaking this is a difficult problem, a point we illustrate with an

[2]The re-weighting problem is similar in its fundamentals to Borusyak and Hull (2023). However, their results cannot be used in our setting because we do not observe the endogenous regressor and their identifying moment assumption is only satisfied asymptotically at a rate that depends on the experimental design.

initial admissibility result: *any* pairing of a clustered design and a Hajak estimator with equal radius performs better, in mean absolute deviation terms, than any alternative pairing of a clustered design and IPW or regression-based estimator for at least one set of potential outcomes. This implies that choosing a particular design and estimator requires either (data-informed) priors or a decision criterion for choosing a design-estimator pair that does not require knowledge of the data-generating process. We consider minimizing the worst-case risk, defining risk as the mean absolute deviation and taking the worst case over all DGPs satisfying mild, plausible conditions.[3] We provide a result that allows the researcher to compute the worst-case risk for any (design, estimator) pair, and thus to choose a pair that minimizes it.

Finally, we provide asymptotically valid inference methods for both the IPW and regression-based estimators. Our confidence intervals are bias-aware and *honest* in the sense of, for example, Armstrong and Kolesár (2020).

2.1.1 Related Literature

We provide a theoretical foundation for and reinterpretation of the literature on program evaluation in the presence of spillovers. At least since Miguel and Kremer (2004) it has been well-understood both that capturing spillover effects is potentially crucial to getting policy inferences right, and that the decay of spillover effects with distance can be exploited to construct estimators that capture them (e.g. Muralidharan et al., 2021; Egger et al., 2022) at least in part (Baird et al., 2016). But it has been unknown whether these estimators are in fact consistent for any estimand of interest, if so under what conditions, or whether they are efficient in any sense given available information. We address each of these questions, providing conditions under which commonly-used estimators are consistent at the optimal rate for the GATE as well as adjustments to improve their performance by attenuating finite-sample bias.

[3]These are bounded outcomes and a requirement that networks not be too connected.

As noted above, our first set of results in Section 2.3 builds on and extends recent theoretical work by Leung (2022b), who studies the case where units are located in $\mathbb{R}^2$. Under the assumption that spillovers decay faster than geometric rate $\eta = 1$ he provides a Horvitz-Thompson estimator that is consistent for the GATE and an experimental design based on growing squares that is rate-optimal over all cluster-randomized designs. We generalize this result to a larger class of spaces, including both arbitrary-dimensional Euclidean spaces as well as non-Euclidean spaces such as directed networks relevant for studying inter-firm dynamics; extend it by characterizing the optimal rate over all linear estimators as well as all clustered designs; and weaken the key decay assumption by allowing for arbitrary rates of decay η. The remainder of the paper then considers performance gains from alternative estimators including those currently used in practice.

Other recent work has focused on estimands other than the GATE, such as the average causal effect of receiving some *specific* level of neighborhood exposure to treatment (Aronow and Samii, 2017; Leung, 2022a). The estimators studied in these papers have some attractive properties (e.g. they are consistent for that estimand even when the exposure mapping is mildly mispecified (Zhang, 2023; Sävje, 2023)), but are not consistent for the GATE since they require designs in which the the probability of universal exposure does not vanish as the population grows. More generally, Basse and Airoldi (2018) show that there is no consistent estimator of the GATE under the assumptions made in these papers.

Finally, the results we obtain under linearity contribute to the literature on linear models of peer effects. These usually specify that outcomes are linear in exogenous and endogenous variables and that the coefficients are known to the researcher up to scale (Manski, 1993; Lee, 2007; Goldsmith-Pinkham and Imbens, 2013; Bramoullé et al., 2014). Others use exposure mappings which rule out any treatment effects by faraway units (Jagadeesan et al., 2020; Sussman and Airoldi, 2017). We show how one can consistently and rate-optimally estimate global average treatment effects without such

assumptions.

2.2 Setup

Notational conventions are as follows. We denote vectors in bold Latin letters $\mathbf{Y}$, fixed deterministic scalars as lowercase Latin or Greek letters n, p, and matrices as capital letters A. A subscripted vector b_i denotes the ith element while a subscripted matrix A_i denotes the ith row. A letter with round brackets denotes a scalar-valued function $\theta(\cdot)$. A script letter $\mathscr{N}$ denotes a set. A script letter with arguments in round brackets $\mathscr{N}(\cdot)$ denotes a set-valued mapping. Fixed universal constants that do not change across populations are also denoted with the letter K subscripted by the number of the assumption that introduced them, e.g. Assumption 3 introduces K_3. A K subscripted by a letter is not linked to an assumption. Where there is little risk of confusion, we suppress subscripts notating dependence on n.

A researcher observes a finite population $\mathscr{N}_n = \{1, 2, \cdots, n\}$ of n units indexed by i. The researcher assigns an $n \times 1$ random vector of treatments $\{d_i\}_{i=1}^n \in \{0,1\}^n$ and observes realized unit outcomes $\{Y_i\}_{i=1}^n \in \mathbb{R}^n$. The potential outcome for household i given the entire treatment assignment is denoted $Y_i(\mathbf{d})$ where the argument is the full $n \times 1$ vector of realized treatment assignments $\mathbf{d}$. Since our estimand is the causal effect of treating an entire population and not a random sample, we model potential outcomes as deterministic functions and the treatment assignment $\mathbf{d}$ as random (and correspondingly will study asymptotics with respect to a growing population).[4] We remain largely agnostic about the origins of the potential outcomes $Y_i(\cdot)$, which could be the outcome of nearly any spatial process, and in particular we do not require that the spillover effects are sparse.

Our estimand of interest is the average causal effect of assigning *every* unit in the

[4]Fixing potential outcomes is standard in this literature, see for example Aronow and Samii (2017); Sussman and Airoldi (2017); Leung (2022a,b).

entire population to treatment vs assigning *no* units at all to treatment. This is called the *global average treatment effect* (GATE) because it sums both the own-effects and spillover effects of treating all units (within some relevant class).

Definition 1. Global Average Treatment Effect (GATE)

$$\theta_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{1}) - Y_i(\mathbf{0})$$

The GATE is often the estimand most relevant for policy decisions; a policymaker using RCT data to decide whether to scale up a program to the control group, for example, would base this decision on the GATE rather than the ATT. Note that we assume the researcher observes outcomes for the entire population for whom she wishes to estimate the GATE.[5] Results concerning the limiting properties of our estimators will require a notion of a growing population, and our estimand θ_n can change as the population grows. Later assumptions will characterize precisely what conditions the sequence of populations must follow.

To solve the core identification challenge we consider restrictions on the *magnitude* of spillovers, which are economically plausible, rather than on their *support*. Specifically, we consider a restriction on the (maximum) impact that faraway treatments have on each unit. Let $\rho_n : \mathscr{N}_n \times \mathscr{N}_n \to \mathbb{R}$ measure the "distance" between any two units. We require $\rho_n(i,i) = 0$ and $\rho_n(i,j) > 0 \, \forall i \neq j$, but place no further restrictions on ρ_n: we do not require symmetry ($\rho(i,j) = \rho(j,i)$) or that the triangle inequality hold exactly, for example. In mathematical terms, ρ_n is a "premetric" but need not be a metric or semimetric. As we discuss below, this will allow us to capture the structure of economic models in which the terms that govern the strength of interaction between units need not satisfy the requirements for a metric (even if the units themselves are located in

[5] We expect the extension to to settings where the study population is a subset of the entire population to be straightforward, but defer this to a subsequent version of the paper.

Euclidean space).

The distance function induces a natural notion of neighborhoods as s-balls:

$$\mathscr{N}_n(i,s) \equiv \{j \in \mathscr{N}_n : \rho_n(i,j) \leq s\} \tag{2.1}$$

Here $s > 0$ is a real number which parameterizes the radius of the neighborhood. This will allow us to study estimators that make use of increasingly large neighborhoods as n grows. Note that every unit is a member of its own neighborhoods ($i \in \mathscr{N}_n(i,s)$) and that for any given n, i there is an $s > 0$ such that $\mathscr{N}_n(i,s) = \{i\}$. Note also that while it is intuitive to think of neighborhoods as being generated in this way, our results hold under *any* definition of neighborhoods that satisfies these two properties, whether they were induced by a distance function or not.

Our first assumption captures the idea of spatial decay in the spillover effects. Heuristically, it says that when s is large and when all of the units in $\mathscr{N}_n(i,s)$ are (non)treated, then unit i's outcome is close to its potential outcome under *universal* (non)treatment. Formally, call a neighborhood "saturated" when all of its members are treated and "dissaturated" when none of its members are treated. Define the random variable $\tilde{s}_i$ as the largest s for which i's s-neighborhood is fully saturated or fully dissaturated.

Definition 2. Largest (Dis)Saturated Neighborhood

$$\tilde{s}_i \equiv \max\{s > 0\} \; s.t. \; d_j = 1 \, \forall \, j \in \mathscr{N}_n(i,s) \; or \; d_j = 0 \, \forall \, j \in \mathscr{N}_n(i,s)$$

Because $i \in \mathscr{N}_n(i,s)$ the largest such neighborhood will be (dis)saturated if i is assigned to (non)treatment. Assumption 6 then states that the difference between unit i's realized outcome and its potential outcome under universal (non)treatment decays geometrically with $\tilde{s}_i$.

Assumption 6. Near Epoch Dependence

There exist universal constants $\eta, K_6 > 0$ *such that:*

$$|Y_i(\mathbf{d}) - d_i Y_i(\mathbf{1}) - (1-d_i)Y_i(\mathbf{0})| \leq K_6 \tilde{s}_i^{-\eta}, \qquad \forall n \in \mathbb{N},\, i \in \mathscr{N}_n,\, \mathbf{d} \in \{0,1\}^n$$

We assume geometric decay as this is consistent in a broad sense with the large body of applied general equilibrium modelling in the "gravity" and "market access" traditions, and mild relative to alternatives such as exponential decay. Faster rates of convergence might of course be achievable under exponential decay, which could also be of interest to study to the extent it can be given micro-economic foundations. Geometric decay will also simplify the statement of results and (we will see) help to make salient the relationship between the rate at which spillovers decay spatially and the rate at which estimators of the GATE converge asymptotically. We do not require any particular geometric rate η, however.

Remark 9. Assumption 6 is closely related to Assumption 3 of Leung (2022b). It is weaker in two ways. First, we bound only interference outside of (dis)saturated neighborhoods while Leung bounds all interference. Second, we do not require that $\eta > 1$.[6] We are able to relax this rate bound by showing that (while the pairwise spillover effects between any two units must indeed decay faster than the square of s) the *sum* of all outside-neighborhood spillover effects can be allowed to decay at *any* geometric rate.

Remark 10. In the language of Definition 1 of Jenish and Prucha (2012), we could say that Assumption 6 guarantees that "Y_i is uniformly L_1-NED with respect to d_i with size η."

Remark 11 (Pairwise Bounds on Interference)**.** An alternative way of embodying the idea that spillovers decay, which might at first seem more conservative, is to assume that

[6]Note that $\eta > 1$ in our notation is equivalent to $\eta > 2$ in Leung's.

the (absolute value of the) effect on unit i of changing a *single* treatment assignment outside unit i's s-neighborhood is less than $s^{-\gamma}$ for some $\gamma > 0$. This assumption turns out to be *stronger* than Assumption 6, however, provided the geography of the neighborhoods is well-behaved. Formally, suppose that for any $s > 0$ and two treatment vectors $\mathbf{d}, \mathbf{d}'$ that differ only at component j where $j \notin \mathscr{N}_n(i,s)$, we have $|Y_i(\mathbf{d}) - Y_i(\mathbf{d}')| < K_0 s^{-\gamma}$. If the maximum size of any s-neighborhood is upper bounded by s^{ω} and $\gamma > \omega > 0$, then Assumption 6 holds with $K_6 = K_0 \frac{\omega}{\gamma - \omega}$ and $\eta = \gamma/\omega - 1$. Proof: Section 2.11.1. The spatial model in Equation 1.3.6 of Kelejian and Piras (2017) satisfies these conditions with $\gamma = 2$ and $\omega = 2$. In addition, the spatial moving average model in Equation (3) of Leung (2022b) satisfies these conditions with his γ equal to our γ and $\omega = 2$.

Assumption 6 ensures that spillover effects decay with space. If in addition the space between units were itself to increase too rapidly as the population grows, then spillovers would become asymptotically irrelevant to the GATE. To focus attention on (more realistic) cases where spillovers are a meaningful challenge, we next introduce an assumption that prevents the population from becoming unboundedly sparse.

To do so we use the concept of an *s-covering* of a set of units $Q \subseteq \mathscr{N}_n$: this is a subset $Q' \subseteq Q$ such that for each $i \in Q$, there is a unit $q \in Q'$ such that $i \in \mathscr{N}_n(q,s)$. The *s-covering number* of Q is then the number of units in the smallest possible s-covering of Q. This conception of the s-covering number always exists. Using this concept we can state Assumption 7.

Assumption 7. Covering Number

There exists a universal constant $K_7 > 0$ such that, for any $Q \subseteq \mathscr{N}_n$ and for all $s > 0$ and all $n \in \mathbb{N}$, the s-covering number of Q is no greater than $K_7 \frac{n}{s}$.

Assumption 7 means that it takes no more than $K_7 \frac{n}{s}$ neighborhoods of members of a subset to cover the entire subset. This guarantees that units are not so sparse that spillovers cease to matter asymptotically. If instead we studied a sequence of

populations where the units rapidly grow farther and farther apart, so that Assumption 7 was violated, the GATE could converge to the ATE, making the problem trivial. Example 3 below illustrates that Assumption 7 holds in familiar Euclidean applications.

Our next assumption controls the distribution of the vector of treatment assignments **d**. Define an "experimental design" as a mapping $\mathscr{D} : \mathscr{N}_n \times \rho_n(\cdot,\cdot) \to F(\mathbf{d})$ from any pair of a population and distance function to a distribution function for the vector of treatments **d**. For a familiar example, the Bernoulli design maps any population and distance function into the CDF of n i.i.d. Bernoulli random variables. However, i.i.d. randomization is unlikely to perform well, since (by the law of large numbers) it tends to generate few large neighborhoods in which most units are treated (or untreated) and which thus approximate the counterfactual scenarios of interest. This motivates examination of a larger class off *cluster-randomized* designs:

Assumption 8. Cluster-Randomized Design

For all populations $\mathscr{N}_n$ and distance functions $\rho_n(\cdot,\cdot)$, $\mathbf{d} \sim \mathscr{D}(\mathscr{N}_n, \rho_n(\cdot,\cdot))$ and:

$$d_i \sim Bern(p)$$

In addition, there exists a partition of $\mathscr{N}_n$ consisting of $c_n \in \mathbb{N}$ clusters called $\{\mathscr{P}_c\}_{c=1}^{c_n}$ such that:

$$d_i = d_j \text{ a.s. if } i,j \in \mathscr{P}_c \text{ for some } c$$

$$d_i, d_j \text{ independent otherwise}$$

Let $c(i)$ denote the cluster that contains unit i.

This assumption allows for a wide variety of cluster-randomized designs used in practice, including those in recent large-scale experimental work. It rules out stratification which (intuitively) we expect to be unnattractive for GATE estimation. Stratification

can create problems for identification because it ensures that the units within a given stratum are never all (un)treated, which in turn means that neighorhoods containing such a stratum can never be entirely (un)treated.[7]

Our final assumption is a regularity condition common in the literature that uniformly bounds potential outcomes. While less restrictive assumptions could be sufficient to prove our results, Assumption 9 eases exposition and is standard (e.g. Aronow and Samii, 2017; Manta et al., 2022; Leung, 2022b).

Assumption 9. Bounded Outcomes

$$|Y_i(\mathbf{d})| < \overline{Y} < \infty, \qquad \forall\, n \in \mathbb{N}, i \in \mathscr{N}_n, \mathbf{d} \in \{0,1\}^n$$

2.3 Rate-Optimal Linear Estimation

We study estimators that are linear, i.e. that take the form of weighted averages of observed outcomes Y_i where the weights $\omega_i(\mathbf{d})$ are bounded functions of the full vector of treatment assignments $\mathbf{d}$. Such estimators can thus be written as

$$\hat{\theta}_n \equiv \sum_{i=1}^{n} Y_i \omega_i(\mathbf{d}), \qquad \sup_{i,n,\mathbf{d}} |n w_i(\mathbf{d})| < \infty \tag{2.2}$$

The IPW estimators commonly used in theoretical analysis take this form, as do most regression-based estimators used in applied work—and all of the estimators studied in this paper (with exception of one which we study in Theorem 9, where we consider a

[7] Notice the absence of an overlap (or "positivity") assumption bounding away from zero the probability that any given unit experiences full exposure to treatment or to control conditions in some relevant neighborhood. Such restrictions on the design are often used when the estimand compares the effects of "local" exposures. In these settings it (helpfully) ensures that the probability that the researcher observes both units whose relevant neighborhoods are fully treated and units whose relevant neighborhoods are fully untreated goes to one, and that the researcher can construct consistent estimators by comparing these unit. In our setting, however, the relevant neighborhood is the entire population, so we can *never* observe both such units at once: if one unit's is exposed to universal treatment, then there cannot be a comparison unit that was exposed to universal non-treatment. As a result the positivity assumption would hurt rather than help us, and we will need to identify our estimand by other means.

special circumstance where the researcher is extremely limited in their choice of design).

We can now state our first main result.

Theorem 5. Optimal Geometric Rate

Let Assumptions 6 and 9 hold. Consider any sequence of populations with distance functions ρ_n satisfying Assumption 7, any sequence of estimators $\hat{\theta}_n$ of the form in Equation 2.2, and any sequence of experimental designs satisfying Assumption 8. Then for every $\alpha, \varepsilon > 0$ there exists $\delta > 0$ such that:

$$\limsup_{n\to\infty} \sup_{\mathbf{Y}\in\mathscr{Y}_n} P\left(n^{\frac{1}{2+\frac{1}{\eta}}+\alpha} \left|\hat{\theta}_n - \theta_n\right| > \varepsilon \right) > \delta$$

Proof: Section 2.11.2

This result states that no estimator taking the form of a weighted mean of outcomes as in (2.2) and no cluster-randomized experimental design can be guaranteed to converge to θ_n at a geometric rate that is strictly faster than $n^{-\frac{1}{2+1/\eta}}$ for all potential outcomes satisfying Assumptions 6 and 9.[8] Notice that the result applies to *any* set of neighborhoods satisfying Assumption 7, and thus is not driven by any pathological configuration of units in space. As the proof illustrates, the potential outcomes that guarantee this rate are also not particularly pathological: they are simply unweighted means of the treatment statuses of nearby neighbors.

When $\eta < \infty$ the convergence allowed by Theorem 5 is slower than the parametric rate $n^{-\frac{1}{2}}$. The result thus illustrates a sense in which GATE estimation in the presence of spillovers with arbitrary support is a more difficult problem than estimation with localized spillovers where the $n^{1/2}$ rate can be guaranteed(Aronow and Samii, 2017). It also links the spatial rate of decay of the spillover effects with the asymptotic rate of convergence of linear estimators. The economic principle this suggests is that estimators based on Assumption 6 will tend to converge faster in less integrated economies where

[8]Additional $\log(n)$ terms are possible.

distance is an important constraint on economic interactions, and converge slower in highly integrated ones where all units are "close" in some sense to each other.

The result builds on Theorem 2 of Leung (2022b) and extends it in several ways. First, we generalize from the case in which the population is contained in a circle in $\mathbb{R}^2$ to the more general set of spaces defined by Assumption 7, which nests $\mathbb{R}^q$ among other spaces. Second, we find the optimal rate over all linear estimators, as well as over all clustered designs. This is of particular interest given that applied work generally uses linear estimators *other* than IPW. Third, we relax the requirement that spillovers (are bounded by a function which) decays at the specific geometric rate $\eta = -1$ and instead accommodate arbitrary geometric decay.

2.3.1 Achieving the optimal geometric rate

We next show that the optimal rate is achievable provided that neighborhoods and population structure are sufficiently "well-behaved." This requires selecting a suitable experimental design, and we will in fact go a bit further and define a class of "Scaling Clusters" designs all of which will permit rate-optimal estimation. This will be useful since many standard designs used in real-world experiments can be interpreted as members of this class.

Clusters must grow with the population size, as will the neighborhoods we use to construct estimators. A fixed cluster size design combined with growing neighborhoods would cause the mean treatment status in any neighborhood to converge rapidly in probability, eliminating the variation of interest. We therefore study clusters that grow with n, governing their rate of growth using an increasing sequence of positive numbers $\{g_n\}_{n=1}^{\infty} \subseteq \mathbb{R}^+$ as follows:

Definition 3. Scaling Clusters Designs

Consider any increasing sequence $\{g_n\}_{n=1}^{\infty} \subseteq \mathbb{R}^+$. A cluster-randomized design is called a "Scaling Clusters" design when there is a $K_D > 1$ such that for every cluster $\mathscr{P}_c$

there is a corresponding unit $q_c \in \mathscr{N}_n$ such that (i) $q_c \in \mathscr{P}_c \subseteq \mathscr{N}_n(q_c, g_n K_D)$ and (ii) the set of $\{q_c\}$ form a g_n-packing of the population.

Condition (ii) refers to the standard concept of an *s-packing*, i.e. a subset of the population such that for all i, j in the packing, $\rho(i,j), \rho(j,i) \geq s$, and is straightforward to verify in practice. To see more transparently that this will require clusters to grow, one can show that (under the assumptions stated later in this section) condition (ii) is satisfied by any design where each cluster contains the neighborhood $\mathscr{N}_n(q_c, K_{10}^{-1} K g_n)$. One can thus think of the Scaling Clusters class as including any design in which clusters both contain and are contained by neighborhoods that grow in proportion to g_n.

In Euclidean spaces it is straightforward to show that a Scaling Clusters design achieves the optimal rate when paired with an appropriate estimator and provided population dynamics are suitable (see Example 3 below for related discussion). Our next assumption defines a class of spaces that are "close enough" to Euclidean for this result to generalize—specifically, loosening the triangle inequality and symmetry properties. It states that if there is a neighborhood that contains both i and j, then the distance between them is bounded by an amount that does not depend on i, j, n, or the absolute size of the neighborhoods.

Assumption 10. Overlap

There is a universal constant $K_{10} \geq 1$, such that

$$i, j \in \mathscr{N}_n(k,s) \implies \rho(i,j), \rho(j,i) \leq K_{10} s, \qquad \forall\, i, j, k \in \mathscr{N}_n, \forall\, n \in \mathbb{N}, \forall s > 0$$

Slight weakenings of this Assumption (as well as Assumption 11 below) are possible, but as these weaker versions become quite technical we state here versions which are relatively easy to check in application. Assumption 10 is automatically satisfied, for example, in any metric space (due to symmetry and the triangle inequality) with $K_{10} =$

2. Allowing for a wider range of K_{10}s allows us to accommodate settings where distances are not metrics, while maintaining a link between coverings and packings. Example 5 below illustrates one case in which this is useful. On the other hand, Assumption 10 rules out neighborhoods that are totally random: it will clearly be violated as the population grows if each distance $\rho(i,j)$ is an i.i.d. random number, for example.

Assumption 10 ensures that a Scaling Clusters design always exists no matter how abstract the space. Algorithm 1 demonstrates how to find one. The algorithm first constructs a g_n-packing of cluster centers called Q. Then the algorithm converts the $K_D g_n$-neighborhoods of each element of Q into clusters until the entire population is assigned to clusters. Since Assumption 10 guarantees that the neighborhoods of the elements of Q are a $K_D g_n$ cover, this algorithm is guaranteed to assign every unit to exactly one cluster.

Algorithm 1. Creating a Scaling Clusters Design

Require: $K_D \geq K_{10}, n \geq 0$, population $\mathscr{N}_n$
  Let $Q = \{x\}$ where $x \in \mathscr{N}_n$
  while $\exists q \in \mathscr{N}_n$ such that $\rho(q',q), \rho(q,q') > g_n \, \forall_{q' \in Q}$ **do**
    Add q to Q
  end while ▷ Q is a g_n packing and a $K_D g_n$ covering
  while $\exists q \in Q$ such that q is not assigned to any cluster **do**
    Create a new cluster consisting of $q \cup (\mathscr{N}_n(q, K_D g_n) \setminus Q)$ minus any units already in a cluster.
  end while

The final assumption we will need to achieve the optimal rate controls population dispersion as n grows, ensuring that the membership of neighborhoods scales (approximately) linearly in s.

Assumption 11. Bounded Density

There exist universal constants $\overline{K}_{11} \geq \underline{K}_{11} > 0$ such that:

$$\underline{K}_{11} s + 1 \leq |\mathscr{N}_n(i,s)| \leq \overline{K}_{11} s + 1 \qquad \forall \, s > 0, \; n \in \mathbb{N}, \, i, j \in \mathscr{N}_n$$

For fixed n one can of course always pick constants such that this holds; what gives it meaning is the restrictions it imposes as n grows. Intuitively one can think of the upper bound as saying that increasing the population means "growing the map" and not solely adding units to a map of a fixed size, in which case it would not be possible to have both saturated and dissaturated neighborhoods of increasing size. The lower bound controls the number of clusters that a Scaling Clusters design can produce by ruling out large numbers of "isolated" units.

Discussion: notions of space

With all the key spatial assumptions (i.e. 7, 10 and 11) now stated, we can take stock of the breadth of application they let us address. Example 3 first verifies they are consistent with standard Euclidean applications:

Example 3. Spatial Designs

Suppose that units are all located in m-dimensional Euclidean space. Let $\rho(i,j) = ||i-j||_2^{1/m}$. *Since this is a metric, Assumption 10 is met immediately with* $K_{10} = 2$. *If no two units can be closer than* ε *together and the entire population is contained within a 2-ball of radius* $n^{1/m}$, *Assumption 11 is also met. Finally, Assumption 7 is met with* $K_7 = 16^m$.[9]

As a simple example of a case in which the flexibility to extend results to non-Euclidean spaces is useful, consider a setting in which there are multiple channels over which spillovers might propagate. For example, training small business owners might lead to business knowledge spillovers within a social network (for which the social network defines the relevant notion of distance), business stealing spillovers for competing firms in the same industry (for which physical distance and product

[9]To see this final point, note that the packing number of a 2-ball of radius $n^{1/m}$ in balls of radius $s^{1/m}$ is upper bounded by $\left(1+4\left(\frac{n}{s}\right)^{1/m}\right)^m$. The δ-covering number of a subset of a 2-ball is no greater than its $\delta/2$-packing number. So the covering number is upper bounded by $\left(1+8\left(\frac{n}{s}\right)^{1/m}\right)^m$. If we restrict $s \leq 2n$, then $8\left(\frac{n}{s}\right)^{1/m} \geq 16$. So the covering number is upper bounded by $16^m\left(\frac{n}{s}\right)$ and so $K_7 = 16^m$. This constant can be sharpened.

substituability define distance), input demand effects on suppliers (for which the buyer-supplier network defines distance), and so on.[10] Which notion of distance should a researcher use? Fortunately they need not choose; our spatial assumptions are flexible enough that they could (for example) work with the minimum across all these metrics, which preserves the essential properties they need:

Example 4. Multiple Distance Notions

Suppose that we have two distance functions ρ_1, ρ_2 and a universal constant C such that $\rho_1(i,j) \leq C\rho_2(i,j)$ and vice versa. If ρ_1, ρ_2 satisfy Assumptions 7, 10, and 11, then $\rho_3(i,j) = \min\{\rho_1(i,j), \rho_2(i,j)\}$ does as well (possibly with different constants). Notice that ρ_3 does not necessarily follow the triangle inequality, even if ρ_1 and ρ_2 do. Proof: Section 2.11.4.

As a fully micro-founded example, we develop in Appendix 2.14 a simple spatial general equilibrium model of a small open economy, and consider the problem of estimating the effects of transfers from a foreign source on total output in the economy. (This setup is designed to roughly parallel the empirical setting and research question in Egger et al. (2022), which we use to discipline our simulations below.) The following example outlines the main ideas, with full details in the appendix.

Example 5. A Gravity Model

Consider a region composed of n locations, each of which produces a distinct, tradeable service using a production function that is linear in labor. Consumers derive utility from a Cobb-Douglass aggregate of their consumption of each of these services, of a numeraire good imported from outside the region, and leisure. Trade between locations incurs iceberg trade costs such that a share τ_{ij} of good shipped from i reaches j; these are related to the

[10] We thank David McKenzie for suggesting this example.

geographic distance $d(i,j)$ between units according to

$$\tau_{ij} \leq \frac{1}{d(i,j)^{\gamma}} \tag{2.3}$$

with $\gamma > 2$. Each household receives a transfer T_i denomoinated in units of the numeraire good. For reasonable parameter values one can show that

1. *Equilibrium trade flows X_{ij} between locations have the "gravity" formulation*

$$X_{ij} \propto \tau_{ij} Y_i Y_j \tag{2.4}$$

   where Y_i is total output in location i;

2. *Spillover effects decay geometrically with distance (i.e. Assumption 6 holds), if we define the distance between two units as*

$$\rho(i,j) = e_{ij} \cdot d(i,j)^2 \tag{2.5}$$

   where e_{ij} is an aggregate of economic parameters of the model.

3. *The distance measure $\rho(i,j)$ as defined in (2.5) satisfies Assumptions 7, 10 and 11;*

4. *Potential outcomes are linear functions of the vector of treatment assignments (i.e. Assumption 12 holds); there exists an exogenous $n \times n$ matrix A such that the vector of location-specific household expenditures $E = (E_1, \ldots, E_n)$ satisfies*

$$E = A(T+1) \tag{2.6}$$

The salient feature of this example is that the measure of distance $\rho(i,j)$ under which Assumption 6 is satisfied is not a proper metric, even though it is influenced by

the Euclidean metric $d(i,j)$. This is because both economics and geography matter in the model. If location i's population is larger than location j's, for example, then ceteris paribus e_{ij} will be greater than e_{ji}, so that $\rho(i,j)$ is asymmetric even if $d(i,j)$ is not. $\rho(i,j)$ does satisfy Assumptions 7, 10 and 11, however, and hence the results developed here are available for studying this economy. Interestingly, potential outcomes are linear in treatment assignment, so that the methods we characterize below in Section 2.4 are also available.

This example considers a sequence of increasingly spatially dense populations. Accommodating a wider range of spatial-economic relationships would be possible if our assumptions were modified in future work.

Consistent estimators

The last step is to define estimators. We consider here the Horvitz-Thompson (HT) and Hajek (HJ) inverse probability weighting (IPW) estimators, which are standard in theoretical work on network effects (see for example Aronow and Samii, 2017; Sussman and Airoldi, 2017; Leung, 2022b). To define them, fix a sequence of neighborhood radii $\kappa_n \subset \mathbb{R}^+$, which will grow to infinity with the sample size. The researcher can choose this sequence, and we will study later how to optimize this choice. Define T_{i1}, T_{i0} as indicators that all of i's neighbors within distance κ_n are treated or untreated:

$$T_{i1} \equiv \prod_{j \in \mathcal{N}_n(i,\kappa_n)} d_j, \qquad T_{i0} \equiv \prod_{j \in \mathcal{N}_n(i,\kappa_n)} (1-d_j)$$

The HT estimator is the weighted difference in outcomes between units with fully treated vs fully untreated neighborhoods.

$$\hat{\theta}^{HT}_{n,\kappa_n} \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i \left(\frac{T_{i1}}{\mathbb{P}(T_{i1}=1)} - \frac{T_{i0}}{\mathbb{P}(T_{i0}=1)} \right) \tag{2.7}$$

While standard, this may perform poorly because the probabilities in the denominators

can be small. It is also sensitive to adding a constant to all outcomes. The HJ estimator makes an adjustment to (partially) address such issues, ensuring that the sum of weights given to units with fully saturated neighborhoods equals the sum of weights given to units with fully dissaturated neighborhoods.

$$\hat{\theta}^{HJ}_{n,\kappa_n} \equiv \sum_{i=1}^{n} Y_i \left(\frac{T_{i1}}{\mathbb{P}(T_{i1}=1)\sum_{k=1}^{n} \frac{T_{k1}}{\mathbb{P}(T_{k1}=1)}} - \frac{T_{i0}}{\mathbb{P}(T_{i0}=1)\sum_{k=1}^{n} \frac{T_{k0}}{\mathbb{P}(T_{k0}=1)}} \right) \tag{2.8}$$

We can now state Theorem 6, which shows that under our assumptions either estimator paired with a Scaling Clusters Design achieves the optimal rate from Theorem 5.

Theorem 6. Inverse Probability Weighting

Let Assumptions 6–11 hold. If the researcher uses a Scaling Clusters Design with $g_n \propto \kappa_n$*, then:*

$$\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n = O_p\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta}\right)$$
$$\hat{\theta}^{HJ}_{n,\kappa_n} - \theta_n = O_p\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta}\right)$$

When $\kappa_n \propto n^{\frac{1}{2\eta+1}}$*, the optimal rate is achieved:*

$$\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta} = O\left(n^{-\frac{1}{2+\frac{1}{\eta}}}\right)$$

Proof: Section 2.11.3.

Unlike the HJ estimator for local effects studied in Aronow and Samii (2017), the HJ estimator of the GATE here converges at a subparametric rate. Unlike the HT estimator in Leung (2022b), our result applies beyond $\mathbb{R}^2$, allows for $\eta \in (0,1)$, and (in conjunction with Theorem 5) shows that the researcher cannot improve on the IPW + Scaling Clusters rate with another *pair* of linear estimator and cluster-randomized design.

2.4 Estimation under linearity

The IPW estimators studied above have not (to our knowledge) been used in practice to estimate the effects of large-scale economic experiments. Instead applied work has typically relied on OLS regressions of outcomes on measures of neighborhood treatment saturation. We turn next to defining and studying the properties of such estimators.

We consider a class of OLS estimators that regress unit-level outcomes on a weighted average of the treatment statuses of nearby clusters. Formally, let $\mathbf{b}$ denote the $c_n \times 1$ vector of i.i.d. Bernoulli(p) cluster treatment assignments and let $\mathbf{v} \equiv \mathbf{b} - p$ be the demeaned cluster treatment assignments. Consider a sequence of $n \times c_n$ matricies B_n of real numbers between 0 and 1. Given such a sequence (and suppressing the subscript from now on) we can define an OLS estimator obtained from a regression of the outcomes on $B\mathbf{v}$ as follows:

$$\hat{\theta}^{OLS}_{n,\kappa_n} \equiv \frac{\widehat{Cov}(B\mathbf{v}, \mathbf{Y})}{\widehat{Cov}(B\mathbf{v}, B\mathbf{v})} \equiv \frac{\frac{1}{n}\mathbf{v}'B'\mathbf{Y} - \overline{B\mathbf{v}}\,\overline{\mathbf{Y}}}{\frac{1}{n}\mathbf{v}'B'B\mathbf{v} - \left(\overline{B\mathbf{v}}\right)^2} \tag{2.9}$$

While it is defined in terms of weighted averages across *clusters*, this class of estimators nests the common specification in which outcomes are regressed on a weighted average of treatment status across *units*. Specifically, setting B to the "unit-weighted" matrix B^U defined by

$$B^U_{ic} = \frac{\sum_{j=1}^{n} \mathbb{1}\left\{j \in \mathscr{N}_n(i, \kappa_n) \cap \mathscr{P}_c\right\}}{|\mathscr{N}_n(i, \kappa_n)|} \tag{2.10}$$

yields the estimator used in Miguel and Kremer (2004), Muralidharan et al. (2021), and Egger et al. (2022). Notice that we interpret the neighborhoods over which treatment status is averaged as growing with n (since κ_n grows), an interpretation not explicitly discussed in the original papers but which admits the possibility of their being consistent for the GATE.

An important idea implicit in regressions of this sort is that treatment effects are linear, in the sense that the effect of treating unit j has the same impact on unit i regardless of any of the other treatment assignments. Assumption 12 formalizes this idea:

Assumption 12. Linear Potential Outcomes

If $\mathbf{d}, \mathbf{d}' \in \{0,1\}^n$ *are disjoint treatment assignment vectors, then:*

$$Y_i(\mathbf{d}+\mathbf{d}') - Y_i(\mathbf{0}) = Y_i(\mathbf{d}) + Y_i(\mathbf{d}') - 2Y_i(\mathbf{0})$$

As an immediate consequence of the Riesz Representation Theorem, for every population $\mathscr{N}_n$, there exists an $n \times n$ matrix A of real numbers with rows denoted A_i such that

$$Y_i(\mathbf{d}) = \beta_0 + A_i\mathbf{d} + \varepsilon_i, \qquad \sum_{i=1}^{n} \varepsilon_i = 0 \tag{2.11}$$

The residual term ε_i is non-random and sums to exactly zero in all populations by construction; since treatment is randomly assigned, $E[\mathbf{d} \mid \varepsilon] = p\mathbf{1}$. The OLS estimator will be consistent no matter how ε is related to the neighborhoods.

While linearity is (arguably) a strong assumption in an economic sense, adding it does not change the optimal rate of convergence that we derived in Theorem 5. The following modification of Theorem 5 states that no estimator that takes the form of a weighted average of outcomes and no cluster-randomized design can guarantee a rate of convergence to θ_n faster than $n^{-\frac{1}{2+\frac{1}{\eta}}}$ when Assumptions 7 and 12 hold.

Theorem 5′. Optimal Rate Under Linearity

Requiring that potential outcomes also satisfy Assumption 12 does not change the conclusions of Theorem 5. Proof: Section 2.11.5.

In other words, the researcher cannot guarantee asymptotic performance any better with linearity than without it. This somewhat surprising result turns out to hinge

on the researchers ability to select any cluster size. We will see below that if the choice of cluster size is constrained then an OLS estimator (which requires the linearity assumption for consistency) may asymptotically outperform other estimators that do not.

With Assumption 12 in hand, we can analyze the convergence of OLS for different choices of B. Theorem 7 provides a convergence result that applies for a broad class of B matrices and for all cluster-randomized designs.

Theorem 7. Convergence of OLS

Let Assumptions 6-12 hold and let the design be a Scaling Clusters design with cluster radius g_n. Let $\kappa_n, g_n \leq n$. Let the $n \times c_n$ matrix B satisfy the following three properties: (i) $B_{ic} \in [0,1]$, (ii) there is a scalar $K < \infty$ such that $B\mathbf{1} \leq K\mathbf{1}$ for all n and (iii) $B_{ic} = 0$ if cluster c is not intersected by $\mathscr{N}_n(i, \kappa_n)$. Then:

$$\frac{\widehat{Cov}(B\mathbf{v}, \mathbf{Y})}{\widehat{Cov}(B\mathbf{v}, B\mathbf{v})} - \frac{tr(B'AP)}{tr(B'B)} = \mathscr{O}_p\left(\frac{\max\{\kappa_n, g_n\}\sqrt{n}}{\sqrt{g_n} tr(B'B)}\right)$$

Proof: Section 2.11.6.

Theorem 7 provides two insights. First, the rate of convergence of OLS depends principally on the column sums of B. In our two examples, the column sums both correspond to how influential each individual's treatment status is on the rest of the population. Second, OLS concentrates around the ratio of traces $\frac{tr(B'AP)}{tr(B'B)}$. This lets us see that the workhorse estimator in the applied literature (i.e. that using B^U) yields a weighted average of treatment effects with weights that are positive but sensitive to the exact specification of the clusters. We call this estimand θ^U. Specifically, some arithmetic yields

$$\frac{tr(B^{U\prime}AP)}{tr(B^{U\prime}B^U)} = \sum_{i=1}^{n}\sum_{k=1}^{n} \underbrace{A_{ik}}_{\text{Effects}} \underbrace{\left(\frac{\frac{\sum_{j=1}^{n} \mathbb{1}\left\{j \in \mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_{c(k)}\right\}}{|\mathscr{N}_n(i,\kappa_n)|}}{\sum_{m=1}^{n}\sum_{c=1}^{c_n}\left(\frac{\sum_{j=1}^{n} \mathbb{1}\{j \in \mathscr{N}_n(m,\kappa_n) \cap \mathscr{P}_c\}}{|\mathscr{N}_n(m,\kappa_n)|}\right)^2}\right)}_{\text{Weights}} \equiv \theta^U$$

Here the causal effect of treating unit k on unit i receives a weight that depends on both k and i. This means that it is not generically feasible to obtain an OLS estimate that is consistent for the unweighted GATE using B^U after first re-weighting the observations. An exception is when randomization is i.i.d. *and* neighborhoods are of uniform size, in which case the weights are uniform.

A consistent OLS estimate of the GATE can be obtained by regressing on the number of treated *clusters* within the neighborhood and dividing by the mean number of clusters intersected by each neighborhood. To do this, set B equal to:

$$B^*_{ic} = \frac{\mathbb{1}\{\mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_c \neq \emptyset\}}{\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{c_n}\mathbb{1}\{\mathscr{P}_{n,c} \cap \mathscr{N}_n(i,s) \neq \emptyset\}} \tag{2.12}$$

This guarantees that the difference between the ratio of traces and the GATE converges to zero at the appropriate rate.

$$\frac{tr(B^{*\prime}AP)}{tr(B^{*\prime}B^*)} - \theta_n = O_p\left(\kappa_n^{-\eta}\right) \tag{2.13}$$

So we can conclude with Theorem 8 which states formally the estimands and rates of convergence when we use the weighting matrix B^*, showing that the latter achieves rate-optimal consistency for the GATE.

Theorem 8. OLS Estimands

If Assumptions 6-12 hold and we use the Scaling Clusters design with scaling sequence g_n *and* $\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\} \Big/ (\sqrt{g_n}\sqrt{n}) \to 0$*, then:*

$$\frac{\widehat{Cov}(B^*\mathbf{b}, \mathbf{Y})}{\widehat{Cov}(B^*\mathbf{b}, B^*\mathbf{b})} - \theta_n = \mathscr{O}_p\left(\frac{\max\left\{\frac{\kappa_n^2}{g_n}, \kappa_n\right\}}{\sqrt{g_n}\sqrt{n}} + \kappa_n^{-\eta}\right)$$

Proof: Section 2.11.7.

Remark 12. *The rate* $\frac{\max\{\kappa_n, g_n\}}{\sqrt{g_n}\sqrt{n}} + \kappa_n^{-\eta}$ *is optimized by setting* $\kappa_n \propto g_n \propto n^{\frac{1}{2\eta+1}}$ *which yields the optimal* $n^{\frac{-1}{2+\frac{1}{\eta}}}$ *rate of convergence from Theorem 5.*

Per Theorem 5′ the rates achieved here are also the best that can be guaranteed under the same assumptions. This result provides some reassurance that the OLS estimators used in earlier studies are close to (though not exactly) consistent and rate-optimal for the GATE.

That said, these results do not tell us whether OLS estimators provide any advantage over HT and HJ given that they are guaranteed to be consistent only under the additional linearity assumption, and yet converge no faster asymptotically. One key differentiating factor here is the degree of flexibility the researcher has to choose the experimental design. Without linearity it is essential that clusters grow quickly, because the only way to learn about the potential outcomes of units surrounded by large (dis)saturated neighborhoods is to actually observe such units. Linearity implies, however, that we can also learn about these potential outcomes from observing units surrounded by large neighborhoods that are only partially (dis)saturated. This means that in a scenario where the researcher cannot grow cluster sizes as fast as she would prefer, so that HT or HJ estimators diverge, she can still obtain consistent estimates assuming linearity and using OLS. Our next theorem illustrates this point by imposing a constraint on the rate of growth of cluster sizes, measured by g_n.

Theorem 9. Small Clusters

If Assumptions 6-12 hold and we use a Scaling Clusters design but set $g_n = n^q$*, where* $q < \frac{1}{2\eta+1}$*, then*

- *The bias of* $\hat{\theta}_{HT}, \hat{\theta}_{HJ}$ *dominates the standard deviation and cannot be guaranteed to decay faster than* $n^{-\eta q}$*.*
- *If we set* $\kappa_n \propto n^{\frac{1+3q}{4+2\eta}}$*, then the bias and standard deviation of* $\hat{\theta}_{OLS}$ *both converge with*

$n^{-\frac{1+3q}{4/\eta+2}}$, *which is faster than* $n^{-\eta q}$.

Proof: Section 2.11.8.

This scenario is practically relevant since in many applied settings researchers do not have total control over the experimental design. In policy experiments, for example, a government may stipulate the administrative level at which the rollout of a reform may be randomized.[11]

Remark 13. *Under certain conditions, OLS can be consistent even without Assumption 12, albeit at a slower rate. Consider for example a setting where units are located on a Euclidean plane and treatment clusters are a tessellation of squares of length* $n^{\frac{1}{4\eta+2}}$. *Then we can show that Assumptions 6–11 are enough to guarantee that OLS with* $\kappa_n = n^{\frac{1}{2\eta+1}}$ *converges at a (slow) rate of* $n^{-\frac{1}{(2+1/\eta)(2+\eta)}}$. *This is a consequence of the geography of this setting and would not necessarily hold if clusters were not tessellations of identical squares. Proof: Section 2.11.9.*

Remark 14. *If the researcher has additional information about the structure of the spillovers, it is possible to construct a shrinkage estimator that exploits this extra information. Suppose that the researcher does not know the true spillover matrix A, but does have a guess* $\hat{A}$. *Appendix 2.10 describes a TSLS estimator where* $B^*\mathbf{d}$ *is used as an instrument for* $A\mathbf{d}$. *When* $\hat{A}$ *meets some minimal conditions, the TSLS estimator is consistent and when the guess is "good enough" in a sense that we specify, the TSLS estimator can show much better small-sample performance than OLS.*

2.5 Admissible and Minimax Design-Estimator Pairs

The convergence results above generally have the following form: if the researcher sets the sizes of the neighborhoods that determine the estimator such that $\kappa_n = C_\kappa n^{\frac{1}{2+\frac{1}{\eta}}}$,

[11] Notice also that when the number of clusters in $\mathcal{N}(i, \kappa_n)$ is increasing, so the bound on the standard deviations in Equation 2.2 does not apply and the assumptions for Theorem 5 do not hold.

and the sizes of randomization clusters such that $g_n = C_g n^{\frac{1}{2+\frac{1}{\eta}}}$, then for a sequence of growing populations the corresponding estimator converges rate-optimally to the GATE. But in practice the researcher must choose κ, g for a given fixed population. We call this the problem of choosing the best "design-estimator-pair" (EDP) for a given experiment. In this section we will provide guidance for choosing a pair of cluster-randomized design and Hajek or Horvitz-Thompson estimator. This guidance may heuristically also apply to the OLS estimator.

2.5.1 Admissibility

Because we impose only weak restrictions on spillover effects, the problem space we must consider when choosing an design-estimator pair is large, making the choice difficult. To illustrate this point we begin by showing that a large number of design-estimator pairs are admissible, even if the researcher knows all of the parameters for all of our assumptions as well as the locations and neighborhoods of every unit in the population. To this end, fix the distance function ρ and the population $\mathscr{N}$. Since we are dealing with a fixed population, suppress all n subscripts for the rest of this section. Let the risk of an design-estimator pair be its expected absolute difference between the estimate and the truth for a given set of potential outcomes. We choose this as our definition of risk because it makes it straightforward to find the risk-maximizing set of potential outcomes given an estimator and cluster-randomized design.

$$\mathscr{R}(\hat{\theta}, \mathscr{D}, \mathbf{Y}(\cdot)) = \mathbb{E}\left[\left|\hat{\theta} - \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}(\mathbf{1}) - \mathbf{Y}(\mathbf{0}))\right| \mid \mathscr{D}, \mathbf{Y}(\cdot)\right] \tag{2.14}$$

where the expectation is taken with respect to random assignments over the distribution induced by the design $\mathscr{D}$. An design-estimator pair $\hat{\theta}, \mathscr{D}$ is said to "dominate" another pair $\hat{\theta}', \mathscr{D}'$ if $R(\hat{\theta}, \mathscr{D}, \mathbf{Y}(\cdot)) \leq R(\hat{\theta}', \mathscr{D}', \mathbf{Y}(\cdot))$ for all potential outcomes $\mathbf{Y}(\cdot)$ satisfying Assumptions 6 and 9, and the inequality is strict for at least one such set of potential

outcomes. An design-estimator pair is "admissible"with respect to some set of alternatives if no other pair within that set dominates it. Now we can state Theorem 10.

Theorem 10. Admissibility of Hajek-SCD Pairs

Let Assumptions 6 and 9 hold. Consider the class of all design-estimator pairs consisting of (i) a Hajak, Horvitz-Thompson, or OLS estimator with radius $\kappa > 0$, and (ii) any cluster-randomized design. Any design-estimator pair consisting of a Hajek estimator and a Scaling Clusters design with equal radius (i.e. $\kappa = g$) is admissible with respect to this class.

Proof: Section 2.12.1

Theorem 10 says that a pairing of a Hajek estimator and a Scaling Clusters design of the same radius is admissible with respect to (essentially) all other pairings of IPW or OLS estimators and cluster-randomized designs. Thus, a large number of design-estimator pairs are justifiable relative to all the other design-estimator pairs considered in this paper. And similar results can be obtained for Horvitz-Thompson results (at the cost of some tedious algebra).

This result is surprising as it shows that even extreme choices like placing all units in a single cluster can be justified. To see why, consider an example where every outcome is zero unless the entire population is treated, in which case every outcome equals a number $M > 0$. We can choose M small enough to satisfy all of our assumptions, even if η is large so that spillovers are known to decay rapidly. Risk is then minimized by placing all units in a single cluster. See the proof of Theorem 10 for detailed examples.

While many design-estimator pairs are admissible, not all of them are. The next example illustrates that pairings of an IPW estimator and a cluster-randomized design in particular may be inadmissible.

Example 6. Not all EDPs are admissible

Consider an experimental design in which $p = 1/2$ and there are C clusters, all of which have at least one member within distance x of every member of the population. Pair this with the Hajek estimator with radius $\kappa > x$. The estimator equals zero with probability $1 - \frac{2}{2^C}$, equals $\overline{\mathbf{Y}(\mathbf{1})}$ with probability $\frac{1}{2^C}$ and equals $-\overline{\mathbf{Y}(\mathbf{0})}$ with probability $\frac{1}{2^C}$, for a risk of $\left|\overline{\mathbf{Y}(\mathbf{1})} - \overline{\mathbf{Y}(\mathbf{0})}\right| \left(1 - \frac{2}{2^C}\right) + \frac{2}{2^C} \left(\left|\overline{\mathbf{Y}(\mathbf{1})}\right| + \left|\overline{\mathbf{Y}(\mathbf{0})}\right|\right)$. So increasing C decreases the risk regardless of the potential outcomes.

2.5.2 Minimizing worst-case risk

In the absence of any additional information, a natural way to select among the large set of admissible design-estimator pairs is to minimize the worst-case-scenario risk. We specify risk here as the expected absolute difference between the estimator and the truth, which has a similar practical interpretation to expected root mean square error and makes maximization tractable. Specifically, we provide below a computable expression for the maximum possible risk of the better of the two of the Horvitz-Thompson and Hajek estimators, given a design $\mathscr{D}$ and a radius κ for the estimator. This allows a researcher to choose those parameters to minimize maximum risk.

To achieve the result, we employ one additional assumption on the structure of the neighborhoods. Assumption 13 says that we can partition the population into a "north" half and a "south" half that have a number of intersections that are small compared to the size of the population.

Assumption 13. North-South

The sequence of populations can be partitioned into two halves N_1, N_2 with equal membership such that for some constant $K_{13} > 0$:

$$\frac{1}{n} \sum_{i \in N_1} \mathbf{1}\{\mathscr{N}_n(i,s) \cap N_2\} + \frac{1}{n} \sum_{i \in N_2} \mathbf{1}\{\mathscr{N}_n(i,s) \cap N_1\} \leq K_{13} s n^{-\frac{1}{2}}$$

This is violated by very connected networks (e.g. "small worlds"), but like

Assumption 10 will typically hold in any application that is similar enough to Euclidean. With this in hand we can now state Theorem 11.

Theorem 11. Maximum Absolute Deviation for IPW

Let Assumptions 6, 7, 11, and 13 hold. Assume also that $|Y_i(\mathbf{1})|, |Y_i(\mathbf{0})| \leq \overline{Y}$ *for all units i. Then for any sequence* $\kappa_n \propto n^{-\frac{1}{2+1/\eta}}$ *and any sequence of Scaling Clusters designs satisfying* $g_n \propto n^{-\frac{1}{2+1/\eta}}$*,*

$$\begin{aligned}
&\max_{\mathbf{Y}} \quad n^{\frac{1}{2+1/\eta}} \min\left\{ \mathscr{R}(\hat{\theta}_n^{HT}, \mathscr{D}, \mathbf{Y}), \mathscr{R}(\hat{\theta}_n^{HJ}, \mathscr{D}, \mathbf{Y}) \right\} \\
&= n^{\frac{1}{2+1/\eta}} \frac{2\overline{Y}}{n} \sqrt{\frac{1}{\pi} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\{\phi_{ij} > 0\} \frac{1}{p^{\phi_{ij}}}} \\
&+ n^{\frac{1}{2+1/\eta}} \frac{K_6}{n} \sum_{i=1}^{n} E\left[\tilde{s}_i^{-\eta} \left(\frac{T_{i1}}{P(T_{i1})} + \frac{T_{i0}}{P(T_{i0})} \right) \right] \\
&+ o(1)
\end{aligned}$$

where ϕ_{ij} *denotes the number of clusters intersected by both* $\mathscr{N}_n(i, \kappa_n)$ *and* $\mathscr{N}_n(j, \kappa_n)$*.*

Proof: Section 2.12.2

The non-vanishing quantities on the right-hand-side depend only on the constants in the assumptions, on the design $\mathscr{D}$ and on the choice of κ_n. The researcher can therefore compute the non-vanishing terms on right-hand-side, e.g. by simulation. Doing so for each of a set of candidate $\kappa, \mathscr{D}$ pairs then allow the researcher to choose a risk-minimizing design. Note that the theorem per se does not obviously imply that the Hajek estimator should be preferred to the Horvitz-Thompson estimator, as the proof demonstrates that they are asymptotically equivalent under at least once sequence of DGP's that aymptotically maximizes the risk of the minimum of the two.

2.6 Bias-Aware Inference

This section presents a method to construct confidence sets covering the GATE. These confidence set are "honest" in the sense that they are bias-aware and achieve adequate coverage without undersmoothing, as in e.g. Armstrong and Kolesár (2020). We focus on ensuring valid coverage without substantially strengthening the assumptions above. Alternative approaches that are simpler to express, or that achieve greater power, at some cost on these first two dimensions may also be possible; a full analysis of these trade-offs is a fruitful direction for further analysis.

There are two principal challenges. To see these, let $\mathscr{F}_i$ denote the σ-algebra generated by the treatment assignments of all units in $\mathscr{N}_n(i,\kappa_n)$. We can then decompose the difference between the GATE and any linear estimator as defined in Equation 2.2 (including all those studied above) in the following way:

$$\sum_{i=1}^{n}\omega_i Y_i-\theta_n=\underbrace{\left(\sum_{i=1}^{n}\omega_i\mathbb{E}[Y_i\,|\,\mathscr{F}_i]-\sum_{i=1}^{n}\mathbb{E}[Y_i\omega_i]\right)}_{[A]}+\underbrace{\sum_{i=1}^{n}\omega_i(Y_i-\mathbb{E}[Y_i\,|\,\mathscr{F}_i])}_{[B]}+\underbrace{\sum_{i=1}^{n}\mathbb{E}[\omega_i Y_i]-\theta_n}_{[C]}$$

Each of these terms will present its own challenges. Term $[A]$ will turn out to be asymptotically normal, but with variance that is not identified. This is well-known consequence of heterogeneous treatment effects in a fixed-population setting (and not a consequence of spillovers or of the choice of estimator per se). Term $[B]$ is zero in expectation and can be interpreted as a random term induced by spillovers across faraway units, as the influence of nearby units have been conditioned out; we cannot control its distribution because of the high amount of dependence across units. Term $[C]$ is a bias term which, when cluster sizes are set to increase at the optimal rate, will converge to zero at the same rate as [A] and so cannot be ignored.

The key decisions are then how to bound the values of the troublesome terms

$[B]$, $[C]$ and the variance of $[A]$. It is possible to achieve both of these goals using only Assumption 6. We suggest, however, using the following strengthening of that assumption, as this yields substantially tighter confidence intervals without much loss of generality.

Assumption 6′. *For every treatment assignment* $\mathbf{d}'$ *and every* $s > 0$*,*

$$\left|Y_i(\mathbf{d}') - E[Y_i(\mathbf{d}) \mid d_i = d_i', \tilde{s}_i \geq s]\right| \leq K_6 \min\{\max\{s^{-\eta}, \tilde{s}_i(\mathbf{d}')^{-\eta}\}, 1\}, \qquad \forall_{i \in \mathcal{N}_n}$$

Assumption 6′ implies Assumption 6 and strengthens it by requiring that (i) the bound on spillover magnitude does not diverge as $\tilde{s}$ gets close to zero, and that (ii) the expectation of Y_i over all the treatment assignments conditional on saturation level s does not fluctuate too much as s grows.The risk-maximizing potential outcomes from Theorem 11 still satisfy Assumption 6′. [12] Assumption 6′ says that potential outcomes settle down quickly as we increase $\tilde{s}$.

Assumption 6′ implies the following bounds on $[B]$ and $[C]$ in terms of observable quantities.

$$|[B]| \leq K_1 \sum_{i=1}^{n} |\omega_i| \min\{\kappa_n^{-\eta}, 1\}$$
$$|[C]| \leq K_1 \sum_{i=1}^{n} E\left[|\omega_i| \min\left\{\tilde{s}_i^{-\eta}, 1\right\}\right]$$

The triangle inequality then yields a bound on the entire difference:

$$\left|\sum_{i=1}^{n} \omega_i Y_i - \theta_n\right| \leq |[A]| + K_6 \sum_{i=1}^{n} E\left[|\omega_i| \min\left\{\tilde{s}_i^{-\eta}, 1\right\}\right] + |\omega_i| \min\{\kappa_n^{-\eta}, 1\} \tag{2.15}$$

This bound will be useful for Horvitz-Thompson or Hajek weights ω_i, but can be very

[12] Assumption 6′ depends on the experimental design; if a researcher wants a condition under which it is guaranteed to hold for any design, the following is sufficient: $|Y_i(\mathbf{d}) - Y_i(\mathbf{d}')| \leq K_6 \min\{\max\{\tilde{s}_i(\mathbf{d})^{-\eta}, \tilde{s}_i(\mathbf{d}')^{-\eta}\}, 1\}, \quad \forall_{i \in \mathcal{N}_n, \mathbf{d}, \mathbf{d}' \in \{0,1\}^n} \implies$ Assum. 6′

conservative in the OLS case because for the OLS weights $\omega_i \neq 0$ does not imply that $\tilde{s}_i$ is large. In the OLS case we instead use the linearity assumption to construct a tighter bound:

Proposition 3. Deviation Bound for OLS. *If Assumptions 6′ and 12 hold,*

$$\left| \sum_{i=1}^{n} \omega_i Y_i - \theta_n \right| \leq |[A]| + K_6 \sum_{i=1}^{n} \min\left\{1, \kappa_n^{-\eta}\right\} \max\{p, 1-p\} \left(|\omega_i| + 1/n\right) \tag{2.16}$$

Proof: Section 2.12.3.

The bounding term on the right hand side is always observed by the researcher. This bound will tend to be loose when κ_n is small, unlike Equation 2.15, but this is arguably not a major concern in practice since (as we have seen) OLS is attractive relative to IPW estimators precisely when κ_n is large compared to the size of the clusters.

To control the asymptotic distribution of $[A]$ we use the following central limit theorem. Let σ_n denote the variance of $n^{\frac{1}{2+1/\eta}}[A]$. Then

Theorem 12. Central Limit Theorem

Let Assumptions 6-11 hold. If the researcher uses a Scaling Clusters design and sets $g_n, \kappa_n \propto n^{\frac{1}{2\eta+1}}$*, and* $\liminf \sigma_n > 0$*, then for the Horvitz-Thompson and Hajek estimators:*

$$\sigma_n^{-1} n^{\frac{1}{2+1/\eta}} \left(\sum_{i=1}^{n} \omega_i E[Y_i \mid \mathscr{F}_i] - \sum_{i=1}^{n} E[Y_i \omega_i] \right) \to_d N(0,1)$$

If in addition Assumption 12 holds, then this result is also true of the OLS estimator. If in addition Assumption 14 holds, this result is also true of the TSLS estimator.

Proof: Section 2.12.4.

This is similar to the CLT in Leung (2022b), stated separately here to match our assumptions and notation. If we knew σ_n^2 it would (in conjunction with Equation 2.15 or 2.16, as appropriate) allow us to construct valid confidence intervals covering the

GATE. The challenge is that σ_n is unidentified. This is true (even in the no-interference and no-treatment-clustering case) due to unobserved heterogeneity of treatment effects in fixed populations. Specifically, the unidentified term in the variance of any linear estimator is

$$\mathscr{R}_n \equiv \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}\left(E[\omega_i Y_i \,|\, \mathscr{F}_i] - E[\hat{\theta}]\right)\left(E[\omega_j Y_j \,|\, \mathscr{F}_i] - E[\hat{\theta}]\right)$$

where Λ denotes the adjacency matrix that connects two units if and only if their κ_n-neighborhoods intersect a common cluster. As this expression illustrates, $\mathscr{R}_n$ is unidentified because we cannot identify the treatment effect of any unit individually. Moreover, its sign is not known, as Λ usually has both positive and negative eigenvalues.

A few approaches to solving this problem have been suggested in the literature. One can always bound $\mathscr{R}_n$ using Young's Inequality (Aronow and Samii, 2017), but this can lead to very conservative confidence sets when treatment clusters are large. An alternative is to require that treatment effect heterogeneity is positively correlated over large enough neighborhoods, in which case $\mathscr{R}_n$ vanishes in large samples (Leung, 2022b). This is attractive when it is defensible. In some of the settings of particular interest to us, in which treatment effects in one large region may be systematically different from those in another. Muralidharan and Niehaus (2017) document, for example, that the effects of the reform studied by Muralidharan et al. (2021) varied substantially across districts in India (see Figure 2). If spillovers are potentially negatively correlated across space, then $\mathscr{R}_n$ could be negative.

We therefore take a new approach, applying Assumption 6′ again to construct a bound on $\mathscr{R}_n$. This yields the following variance estimators (the last three terms of each of which is the bound on $\mathscr{R}_n$). The expressions are admittedly quite involved to state, but the important point is that they can always be computed and are guaranteed to be conservative in large samples. Let Q denote the block-diagonal adjacency matrix that con-

nects two units if and only if they are in the same cluster. We call the Horvitz-Thompson estimator the variance estimator $\widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i\,|\,\mathscr{F}_i]\right)$, the Hajek variance estimator $\widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{HJ}E[Y_i\,|\,\mathscr{F}_i]\right)$, and the OLS variance estimator $\widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{OLS}E[Y_i\,|\,\mathscr{F}_i]\right)$. Statements of the variance estimators is relegated to Appendix 2.9 because they are lengthy. Theorem 13 says that these variance estimators are guaranteed to be conservative for the variance of $[A]$.

Theorem 13. Variance Estimation

If Assumptions 6-11 and 6′ hold and $g_n, \kappa_n \propto n^{\frac{1}{2\eta+1}}$*, then:*

$$\liminf_{n\to\infty} n^{\frac{1}{2+1/\eta}}\left(\widehat{Var}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i\,|\,\mathscr{F}_i]\right)-Var\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i\,|\,\mathscr{F}_i]\right)\right)\geq 0$$
$$\liminf_{n\to\infty} n^{\frac{1}{2+1/\eta}}\left(\widehat{Var}\left(\sum_{i=1}^{n}\omega_i^{HJ}E[Y_i\,|\,\mathscr{F}_i]\right)-Var\left(\sum_{i=1}^{n}\omega_i^{HJ}E[Y_i\,|\,\mathscr{F}_i]\right)\right)\geq 0$$

If in addition Assumption 12 holds,

$$\liminf_{n\to\infty} n^{\frac{1}{2+1/\eta}}\left(\widehat{Var}\left(\sum_{i=1}^{n}\omega_i^{OLS}E[Y_i\,|\,\mathscr{F}_i]\right)-Var\left(\sum_{i=1}^{n}\omega_i^{OLS}E[Y_i\,|\,\mathscr{F}_i]\right)\right)\geq 0$$

Proof: Section 2.12.5.

We can now combine all of the previous results to state the following 95% confidence interval for the Horvitz-Thompson estimator:

$$\begin{aligned}CI_{HT} &\equiv \left[\hat{\theta}^{HT}-\delta_{HT},\hat{\theta}^{HT}+\delta_{HT}\right]\\ \delta_{HT} &\equiv 1.96\sqrt{\widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i\,|\,\mathscr{F}_i]\right)}\\ &\quad + K_6\sum_{i=1}^{n}E\left[|\omega_i|\min\left\{\tilde{s}_i^{-\eta},1\right\}\right]+|\omega_i|\min\{\kappa_n^{-\eta},1\}\end{aligned}$$

The confidence set for the Hajek estimator is identical except for the variance estimator.

For the OLS estimator the radius of the confidence interval is:

$$\delta_{OLS} \equiv 1.96\sqrt{\widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{OLS}E[Y_i\,|\,\mathscr{F}_i]\right)} + K_6\sum_{i=1}^{n}\min\left\{1,\kappa_n^{-\eta}\right\}\max\left\{p,1-p\right\}\left(|\omega_i|+1/n\right)$$

If we combine Theorems 12 and 13 with Equation 2.15 or Proposition 3, then Corollary 3 is immediate. It states that our confidence sets achieve at least 95% coverage asymptotically.

Corollary 3. Inference

If Assumptions 6-11 and 6′ hold and the researcher sets $g_n = \alpha\kappa_n \propto n^{\frac{1}{2\eta+1}}$ *with* $\alpha > 0$

$$\limsup_{n\to\infty} P(\theta \in CI_{HT}) \geq 0.95$$

$$\limsup_{n\to\infty} P(\theta \in CI_{HJ}) \geq 0.95$$

If in addition Assumption 12 holds,

$$\limsup_{n\to\infty} P(\theta \in CI_{OLS}) \geq 0.95$$

2.7 Simulations

This section uses simulations to study the small-sample performance of the Horvitz-Thompson, Hajek and OLS estimators, as measured by the mean absolute difference between the estimate and the truth, as well as the coverage and radius of the corresponding 95% confidence intervals.

The specification of the simulations is as follows. We first assign units uniformly at random to positions within a two-dimensional circle with radius $\sqrt{n}$.[13] We define $\rho(i,j)$ as the square root of the L_2 distance between units i and j. This satisfies Assumptions

[13]We do enforce an minimum Euclidean distance of 0.1 between units by dropping and resampling.

7, 10, and 11. We then generate outcome data using four different data-generating processes (DGPs) selected to contrast the performance of different estimators under different conditions.

The "IID" DGP is a benchmark case with homogeneous treatment effects and no spillovers. It thus trivially satisfies Assumptions 6, 6′, and 12. The bounded (here, uniform) distribution for the error terms ensures that $|Y_i(\mathbf{1})|, |Y_i(\mathbf{0})| \leq 2$, which by Theorem 11 guarantees that DGP 1 will have no more risk in large samples under HT and HJ than the worst-case DGP we will introduce shortly (DGP 4). Note that we draw the residual terms ε once at the start of the simulation and hold them constant over each draw of the treatment assignments.

DGP 1. IID

$$\begin{aligned} Y_i(\mathbf{d}) &= d_i + \varepsilon \\ \varepsilon_i &\overset{\text{iid}}{\sim} U(-1,1) \end{aligned}$$

The "Linear" DGP adds spillovers, spatial autocorrelation in the error terms, and some treatment effect heterogeneity, while maintaining linearity in treatment effects. It too satisfies Assumptions 6, 6′, and 12. In order to ensure again that this DGP has less large-sample risk than DGP 4 we again choose error terms such that $|Y_i(\mathbf{1})|, |Y_i(\mathbf{0})| \leq 2$, here via a normalization.

DGP 2. Linear Spillovers

$$\begin{aligned} Y_i(\mathbf{d}) &= A_i\mathbf{d} + \mu_i \\ A_{i,j} &= \mathbf{1}\{i=j\} + \mathbf{1}\{i \neq j\} K_6 \max\{1, \rho(i,j)\}^{-2(\eta+1)} \\ \mu_i &= \frac{A_i\varepsilon}{max(A_i\mathbf{1})}(2 - A_i\mathbf{1}) \end{aligned}$$

The "Nonlinear" DGP is specified to specifically demonstrate potential drawbacks of OLS with large neighborhoods (large κ_n). Here the global average treatment effect is always zero, but units receive a "relative" effect when their nearby neighborhood is fully treated which diminishes to zero as the width of this neighborhood grows. This is thus a "zero sum" environment when a unit's treatment effect diminishes as more of its neighbors are treated. It satisfies Assumptions 6 and 6′, but not Assumption 12.

DGP 3. Nonlinear Spillovers

$$Y_i(\mathbf{d}) = (2d_i - 1)K_6 \min\left\{1, \tilde{s}_i^{-\eta}\right\} + \varepsilon_i$$

Finally, DGP 4 is the worst-case-scenario DGP that asymptotically maximizes expected absolute deviation for the better of the Horvitz-Thompson and Hajek estimators as in Theorem 11. Because of this (and unlike the other DGPs) it depends on the experimental design and the radius of the Horvitz-Thompson estimator. It serves as a useful benchmark for assessing to what extent (if any) there is a tradeoff between controlling worst-case risk and controlling risk in other more typical scenarios. This DGP satisfies Assumptions 6 and 6′, but not Assumption 12. Moreover, we have $|Y_i(\mathbf{1})|, |Y_i(\mathbf{0})| \leq 2$, as under the previous DGPs; by Theorem 11 this implies that DGP 4 will have the (weakly) greatest HT and HJ risk of the four in large samples. Note, however, that this guarantee holds *only* asymptotically, i.e. for n large relative to κ, g.

DGP 4. IPW Risk Maximizer

$$e_i \equiv 4\mathbf{1}\{i \text{ north of median y-coordinate}\} - 2$$
$$Y_i(\mathbf{d}) = Sign\left(\sum_{j=1}^{n} e_j \omega_j^{HT}\right)(2d_i - 1)K_6 \min\left\{1, \tilde{s}_i^{-\eta}\right\} + e_i$$

After generating potential outcomes once as above, we simulate observed out-

comes for 500 experimental assignments. We assign treatment using a Scaling Clusters design generated using Algorithm 1. We then calculate estimates using the HT, HJ and OLS, estimators. We set $\kappa_n = C_\kappa n^{\frac{1}{2+1/\eta}}$ and $g_n = C_g n^{\frac{1}{2+1/\eta}}$, matching the optimal rate of scaling, while varying the scaling constants.

We first examine how performance varies with sample size. We expect slower-than-parametric rates of convergence for all the estimators; specifically, Theorems 6 and 7 imply that their risk should decay with $n^{-1/3}$. Table 2.1 shows that for all three DGPs, the expected absolute deviation and radius of the confidence intervals shrink as the sample size grows, and that risk decays at a slower-than-parametric rate. If in contrast the estimators were converging at the parametric rate, then increasing n from 250 to 1000 would cut risk in half.

Table 2.1 also confirms the assertions of Theorem 11 that the risk of HT and HJ are the same for the maximizer DGP, and larger than for the other DGPs for all values of κ, g. Thus choosing κ, g to minimize risk of the maximizer DGP does minimize worst-case risk over the DGPs in these simulations, as predicted. Notice also that OLS performs just as well as Hajek even under the nonlinear DGPs; this is because when $g_n = \kappa_n$ in two-dimensional Euclidean space, the two estimators typically nearly agree.

Since we cannot avoid using conservative confidence intervals, coverage is at least 95% for all parameterizations (and is in fact frequently close to 100%). We can quantify how conservative the confidence sets are by comparing them to the expected absolute deviation. If $X \sim N(\mu, \sigma)$, then inverting a t-test yields an exactly 95% confidence interval covering μ that has radius 1.96σ. The expected absolute deviation of X from its true mean is $\sqrt{2/\pi}\sigma$. So the radius of the confidence interval covering μ is $1.96\sqrt{\pi/2} = 2.45$ times larger than its expected absolute deviation. Table 2.1 shows that for the IID and Linear DGPs and $n = 1000$ the confidence interval is about 12-13 times larger than the expected absolute deviation, so these intervals are wide. But for the maximizer DGP this ratio is only 2.59, quite close to the benchmark example—so these confidence sets can

be quite tight.

The second simulation exercise illustrates the strengths and weaknesses of the OLS estimator. For this exercise we force the researcher to randomize with only unit per cluster, setting $g_n = 0$. Recall that Theorem 9 implies that—assuming linearity (Assumption 12) holds—OLS has advantages in such scenarios. The first half of Table 2.2 illustrates this: both risk and the radius of the 95% confidence interval are minimized by using the OLS estimator with a large κ. In the second half of the table, on the other hand, the DGP is nonlinear and the risk of OLS is higher than that of HT or HJ, with the difference particularly large. This is noteworthy since in applications randomization clusters often are in fact small relatively to neighborhood sizes, as for example in Miguel and Kremer (2004); Egger et al. (2022), and researchers use OLS estimators as a way of compensating for this.

2.8 Tables for Chapter 2

Table 2.1. Convergence

				Risk			CI width		
DGP	n	g_n	κ_n	HT	HJ	OLS	HT	HJ	OLS
iid	250	3.00	3.00	0.09	0.02	0.04	0.50	0.39	0.39
iid	500	3.78	3.78	0.08	0.02	0.03	0.44	0.32	0.31
iid	1000	4.76	4.76	0.06	0.02	0.03	0.34	0.26	0.25
Linear	250	3.00	3.00	0.11	0.06	0.06	0.54	0.41	0.39
Linear	500	3.78	3.78	0.10	0.05	0.05	0.48	0.34	0.32
Linear	1000	4.76	4.76	0.09	0.03	0.04	0.42	0.26	0.26
Nonlinear	250	3.00	3.00	0.12	0.11	0.12	0.44	0.39	0.32
Nonlinear	500	3.78	3.78	0.10	0.10	0.10	0.39	0.33	0.27
Nonlinear	1000	4.76	4.76	0.08	0.07	0.08	0.32	0.26	0.22
Maximizer	250	3.00	3.00	0.53	0.53	0.53	1.40	1.38	1.32
Maximizer	500	3.78	3.78	0.48	0.48	0.48	1.21	1.19	1.13
Maximizer	1000	4.76	4.76	0.42	0.42	0.42	1.04	1.03	0.98

Shows the convergence of three estimators over four DGPs. Risk is mean absolute deviation from the AGE. Width of the confidence intervals is displayed in the last three columns. We use $\eta = 1$ and $K_6 = 0.33$ for all parameterizations. The smallest coverage for any parameterization and any estimator was 94.2%.

Table 2.2. Relative performance of OLS and IPW with small clusters

				Risk			CI width		
DGP	n	g_n	κ_n	HT	HJ	OLS	HT	HJ	OLS
Linear	1000	0	0	0.27	0.27	0.27	0.93	0.90	0.69
Linear	1000	0	1	0.24	0.24	0.23	1.04	0.95	0.80
Linear	1000	0	2	0.16	0.11	0.11	0.75	0.55	0.50
Linear	1000	0	3	0.21	0.09	0.08	0.74	0.46	0.42
Linear	1000	0	4	0.27	0.08	0.06	0.83	0.39	0.37
Linear	1000	0	5	0.34	0.09	0.06	0.89	0.37	0.35
Nonlinear	1000	0	0	0.42	0.43	0.43	0.91	0.90	0.69
Nonlinear	1000	0	1	0.33	0.34	0.35	0.98	0.95	0.77
Nonlinear	1000	0	2	0.20	0.20	0.28	0.59	0.54	0.47
Nonlinear	1000	0	3	0.15	0.15	0.25	0.49	0.43	0.37
Nonlinear	1000	0	4	0.13	0.12	0.24	0.46	0.37	0.32
Nonlinear	1000	0	5	0.15	0.10	0.24	0.52	0.37	0.31

Shows advantages and disadvantages of OLS vs IPW when cluster size is constrained. These simulations use $\eta = 1$ and $K_1 = 0.33$. Risk is expected absolute deviation. Coverage of the confidence intervals is larger than 95% in all cases except for the Horvtiz-Thompson CI for the Linear DGP and $\kappa = 5$ for which coverage is 89%.

2.9 Lengthy Statements

This appendix includes statements of quantities that are too lengthy to include in the main text.

For the Horvitz-Thompson estimator the variance estimator is:

$$\begin{aligned}
\widehat{\mathrm{Var}}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i\,|\,\mathscr{F}_i]\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}Y_iY_j\left(\frac{T_{i1}T_{j1}}{P(T_{i1})P(T_{j1})}+\frac{T_{i0}T_{j0}}{P(T_{i0})P(T_{j0})}\right)\\
&\quad-\hat{\theta}_{HT}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}(2\omega_i^{HT}Y_i-\hat{\theta}_{HT})\\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})K_1^2(\tilde{s}_i^{-\eta})(\tilde{s}_j^{-\eta})\\
&\quad-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})\left(\frac{2d_i-1}{p}Y_i-\hat{\theta}_{HT}\right)\left(\frac{2d_j-1}{p}Y_j-\hat{\theta}_{HT}\right)\\
&\quad+\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})\left|\frac{2d_i-1}{p}Y_i-\tilde{\theta}_{HT}\right|K_1\tilde{s}_j^{-\eta}
\end{aligned}$$

For the Hajek estimator:

$$\begin{aligned}
r_{1i} &\equiv Y_i-\frac{1}{n}\sum_{k=1}^{n}\frac{T_{k1}Y_k}{P(T_{k1})}\\
r_{0i} &\equiv Y_i-\frac{1}{n}\sum_{k=1}^{n}\frac{T_{k0}Y_k}{P(T_{k0})}\\
\widehat{\mathrm{Var}}\left(\sum_{i=1}^{n}\omega_i^{HJ}E[Y_i\,|\,\mathscr{F}_i]\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}r_{i1}r_{i0}\left(\frac{T_{i1}T_{i0}}{P(T_{i1})P(T_{j1})}\right)\\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}r_{0i}r_{0j}\left(\frac{T_{i0}T_{j0}}{P(T_{i0})P(T_{j0})}\right)\\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})K_1^2\tilde{s}_i^{-\eta}\tilde{s}_j^{-\eta}\\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})\left(d_ir_{i1}-(1-d_i)r_{i0}\right)\left(d_jr_{j1}-(1-d_j)r_{j0}\right)\\
&\quad+\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})\left(d_ir_{i1}-(1-d_i)r_{i0}\right)K_1\tilde{s}_j^{-\eta}
\end{aligned}$$

For the OLS estimator:

$$
\begin{aligned}
e_i &\equiv x_i \left(y_i - x_i \tilde{\theta}\right) \\
D &\equiv \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^2\right]\right)^2} \\
\widehat{\mathrm{Var}}\left(\sum_{i=1}^{n} \omega_i^{OLS} E[Y_i \mid \mathscr{F}_i]\right) &= \frac{D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Lambda_{ij} e_i e_j \\
&\quad + \frac{D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij}) K_1^2 \tilde{s}_i^{-\eta} \tilde{s}_j^{-\eta} \\
&\quad - \frac{D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij}) \left(\frac{2d_i - 1}{p} Y_i - \hat{\theta}_{OLS}\right)\left(\frac{2d_j - 1}{p} Y_j - \hat{\theta}_{OLS}\right) \\
&\quad + \frac{2D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij}) \left|\frac{2d_i - 1}{p} Y_i - \hat{\theta}_{OLS}\right| K_1 \tilde{s}_j^{-\eta}
\end{aligned}
$$

2.10 A TSLS Estimator

The OLS estimator based on B^* makes no use of any information about the structure of spillovers beyond that contained in the distances $\rho_n(i,j)$. In practice, the researcher may have additional information about the patterns of interaction between units. For example, a researcher studying cash transfers to households might have information on where households shopped or worked prior to the intervention; a researcher studying the effects of export taxes might observe input-output linkages between firms as well as their geographic proximity. A natural question is whether the researcher can use such auxiliary information to improve performance.

We examine here an estimator based on two-stage-least-squares regression, where geographic exposure to treatment is used as an instrument for a "guess" for treatment exposure based on auxiliary information. Using geography as the instrument for the guess (as opposed to the other way around) creates flexibility for the guess because the potential outcomes need not satisfy satisfy Assumption 6 under the weighted distances

it implies. We think of this as a shrinkage estimator as it "shrinks" the OLS estimate towards one implied by the auxiliary information. When the quality of the auxiliary information is high in a sense that we specify, this estimator can have smaller bias than OLS.

We start by making precise what we mean by "auxiliary information." Under linearity (Assumption 12), spillovers can be described in terms of the $n \times n$ weighted adjacency matrix A from Equation 2.11. The researcher does not observe A but does observe another $n \times n$ matrix $\hat{A}$, which we refer to as a "guess" (rather than an estimate) as it need not converge to A and is not random. We require instead only that $\hat{A}$ meet the four requirements described in Assumption 14. First, the guess $\hat{A}$ must be deterministic in the sense that it is unrelated to treatment assignment. For instance, $\hat{A}$ could be a the outcome of a calibrated general equilibrium model that might be misspecified, or it could be based on survey data collected before treatment was assigned. Second and third, the guess must itself satisfy conditions such that if it were the true spillover matrix then our assumptions of geometric decay (Assumption 6) and bounded outcomes (Assumption 9) would hold. These are easy to meet, in the sense that a researcher with a guess that did not satisfy them could always modify that guess to impose them. Finally, the sum of the elements $\hat{A}$ must be bounded away from zero. This last requirement, when combined with the second, will be sufficient for the instrument strength requirement for TSLS regression.

Assumption 14. Requirements for the guess $\hat{A}$

For each population $\mathcal{N}_n$ there is an $n \times n$ matrix $\hat{A}_n$ (for which we will from now on suppress the subscript) which satisfies:

1. **Fixed:** *$\hat{A}$ is deterministic.*

2. **Geometric Interference:** $\sum_{j \notin \mathcal{N}_n(i,s)} |A_{i,j}| \leq K_6 s^{-\eta}$ *for all $i \in \mathcal{N}_n$ and all $s > 0$*

3. **Bounded Outcomes:** $\sum_{j=1}^{n} |A_{i,j}| \leq \overline{Y}$ *for all* $i \in \mathcal{N}_n$

4. **Instrument Strength:** $\frac{|\mathbf{1}'\hat{A}\mathbf{1}|}{n} > K_{14}$ *for some* $K_{14} > 0$

With a valid $\hat{A}$ in hand, the researcher can construct a TSLS estimator for the GATE. If we define a transformation of the outcomes as

$$\tilde{\mathbf{Y}} \equiv \mathbf{Y}\left(\frac{\mathbf{1}'\hat{A}\mathbf{1}}{n}\right) \tag{2.17}$$

we can specify the TSLS estimator as the result of a TSLS regression of the transformed outcome $\tilde{\mathbf{Y}}$ on $\hat{A}P\mathbf{v}$, instrumented with $B^*\mathbf{v}$, where P is an $n \times c_n$ matrix such that P_{ic} indicates whether unit i belongs to cluster c:

$$\hat{\theta}^{TSLS}_{n,\kappa_n} \equiv \frac{\widehat{Cov}\left(B^*\mathbf{v}, \tilde{\mathbf{Y}}\right)}{\widehat{Cov}\left(B^*\mathbf{v}, \hat{A}P\mathbf{v}\right)} \tag{2.18}$$

This estimator is consistent even when the guess $\hat{A}$ is very far from the truth; in fact, for any guess $\hat{A}$ satisfying Assumption 14 it converages at the optimal rate, just as OLS does:

Theorem 14. Convergence of TSLS Estimator

If Assumptions 6-14 hold and we use a Scaling Clusters design and set $g_n, \kappa_n \propto n^{\frac{1}{2\eta+1}}$*, then:*

$$\hat{\theta}^{TSLS}_{n,\kappa_n} - \theta_n = O_p\left(n^{-\frac{1}{2+\frac{1}{\eta}}}\right)$$

Proof: Section 2.12.6.

The choice between TSLS and OLS thus depends on their bias and variance. This comparison is complicated by the fact that TSLS does not have a defined expectation

(as it is a ratio with a random denominator whose support includes zero), so that standard measures of performance such as MSE are not applicable. We therefore make the comparison by examining analytically the limiting amount by which each estimator differs from θ_n when we hold κ_n and g_n constant while increasing n to infinity. This thought experiment corresponds to fixing the research design and estimators actually used by the researcher for a given finite population, and asking how they perform in populations large enough for moments to converge to their limits.

To implement this approach, let the matrix $\tilde{A}(\kappa_n)$ denote the part of A (and analogously $\widehat{\tilde{A}}$ the part of $\hat{A}$) that captures spillovers onto unit i coming from units j in clusters that do not intersect i's κ_n-neighborhood:

$$\tilde{A}_{ij}(\kappa_n) = \begin{cases} 0 & c(j) \cap \mathcal{N}_n(i,\kappa_n) \neq \emptyset \\ A_{ij} & \text{otherwise} \end{cases} \qquad \widehat{\tilde{A}}_{ij}(\kappa_n) = \begin{cases} 0 & c(j) \cap \mathcal{N}_n(i,\kappa_n) \neq \emptyset \\ \hat{A}_{ij} & \text{otherwise} \end{cases}$$

Using this notation, $\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{n}$ denotes the portion of the estimand that is due to spillovers that occur outside of clusters that intersect the κ_n-neighborhoods. Suppressing the κ_n argument for simplicity, we can now state:

Theorem 15. Comparing OLS and TSLS Estimators

Under Assumptions 6-14, if $\kappa_n = \kappa^* > 0$ and $g_n = g^* > 0$ are fixed constants and $\mathbf{1}'\widehat{\tilde{A}}\mathbf{1} \neq \mathbf{1}'\widehat{A}\mathbf{1}$, then:

$$\hat{\theta}^{OLS}_{n,\kappa^*} - \theta_n + \left(\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{n}\right) = O_p\left(n^{-\frac{1}{2}}\right)$$

$$\hat{\theta}^{TSLS}_{n,\kappa^*} - \theta_n + \left(\frac{\mathbf{1}'A\mathbf{1}}{n}\right)\left(\frac{\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{\mathbf{1}'A\mathbf{1}} - \frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}}{1 - \frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}}\right) = O_p\left(n^{-\frac{1}{2}}\right)$$

Proof: Section 2.12.7

Theorem 15 sheds some light on when to use TSLS rather than OLS, and in doing

so on what makes a good guess $\hat{A}$. It reveals that the difference between the estimators depends on how well the guessed fraction of spillovers that are off-neighborhood $\frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}$ approximates the true fraction of spillovers that are off-neighborhood $\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{\mathbf{1}'A\mathbf{1}}$. If the ratio of off-neighborhood spillovers to on-neighborhood spillovers is the same for the guess and for the truth, the "bias" term for the TSLS estimator is zero, making the TSLS estimator preferable in this sense. More generally, if every entry of both matrices is positive, i.e. $A_{ij} > 0$ and $\hat{A}_{ij} > 0$, then one can show that the absolute value of the "bias" of the TSLS estimator is smaller than the absolute value of the "bias" of OLS if and only if:

$$\frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}\left(1+\left(\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{\mathbf{1}'A\mathbf{1}}\right)^{-1}\right) < 2$$

This result can be interpreted in the following intuitive way. OLS "misses" all the off-neighborhood spillovers. The TSLS estimator tries to correct this error by scaling up the OLS estimate by the proportion of spillovers that, according to the guess $\hat{A}$, are off-neighborhood. When this proportion is close to the true proportion, so that $\frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}\left(\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{\mathbf{1}'A\mathbf{1}}\right)^{-1} \simeq 1$, the TSLS estimator thus targets a quantity closer to the truth than OLS.

2.11 Proofs for Chapter 2

2.11.1 Proof of claim in Remark 11

Under the proposition's hypotheses, the quantity $|Y_i(\mathbf{d}) - Y_i(\mathbf{d}')|$ when $\mathbf{d}, \mathbf{d}'$ differ only off of the $\mathscr{N}_n(i,s)$ is upper bounded by the sum over all $x > s$ of $x^{-\gamma}$ times the number of possible members of the x neighborhood but not the $x-1$ neighborhood. Regardless of the population size, this is upper bounded by $\sum_{x=s}^{\infty} x^{-\gamma}(x^{\omega} - (x-1)^{\omega})$. This sum is a lower Riemann sum and is upper bounded by the integral: $\int_s^{\infty} x^{-\gamma/\omega} dx = \frac{\omega s^{1-\gamma/\omega}}{\gamma-\omega}$ which implies Assumption 6 with $K_6 = \frac{\omega}{\gamma-\omega}$ and $\eta = \gamma/\omega - 1$ when $\gamma > \omega > 0$.

2.11.2 Proof of Theorem 5

First define the following notation: Let $\phi_n(i,s)$ denote the number of clusters that intersect the s-neighborhood of unit i. Let $\gamma_n(c,s)$ denote the number of s-neighborhoods that intersect cluster c. Formally,

$$\phi_n(i,s) \equiv \sum_{c=1}^{c_n} \mathbb{1}\{\mathscr{P}_{n,c} \cap \mathscr{N}_n(i,s) \neq \emptyset\}$$
$$\gamma_n(c,s) \equiv \sum_{i=1}^{n} \mathbb{1}\{\mathscr{P}_{n,c} \cap \mathscr{N}_n(i,s) \neq \emptyset\}$$

We omit the i or c argument when we wish to denote the maximum over a population.

$$\phi_n(s) \equiv \max_{i \in \mathscr{N}_n} \phi_n(i,s)$$
$$\gamma_n(s) \equiv \max_{c \in \mathscr{P}_n} \gamma_n(i,s)$$

To prove this theorem, we need this lemma. Its proof is in the next appendix.

Lemma 8. *If $\hat{\theta}_n - \theta_n = O_p(a_n)$ for all potential outcomes allowed under Assumptions 6, 9, and 12 under some sequence of designs satisfying Assumption 8, then for every $\varepsilon \in \left(0, \frac{1}{\eta}\right)$ and every $\delta > 0$*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\} = o(1)$$

Proof: Section 2.13.1.

Here is the proof of Theorem 5.

Consider potential outcomes $Y_i^*(\mathbf{d})$ that is zero for units where $\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}$. This does not violate any of our assumptions. Since $\hat{\theta}_n$ is still consistent at rate a_n:

$$\sum_{i=1}^{n} Y_i^*(\mathbf{d})\omega_i(\mathbf{d}) - \theta_n^* = O_p(a_n)$$

$$\sum_{i=1}^{n} Y_i^*(\mathbf{d})\omega_i(\mathbf{d})\mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) < a_n^{-\delta}\right\} - \theta_n^* = O_p(a_n)$$

We will show that in order for the estimator to be a_n-consistent for these potential outcomes, a_n cannot shrink faster than $n^{-\frac{1}{2+\frac{1}{\eta}}}$.

Now use Assumption 7. Our goal is to upper bound the number of clusters that contain units with nonzero treatment effects. Assumption 7 guarantees that the $a_n^{-\frac{1}{\eta}+\varepsilon}$-covering number of the set of units with nonzero treatment effects is no greater than $K_7 \frac{n}{a_n^{-\frac{1}{\eta}+\varepsilon}}$. By construction, each of the neighborhoods in the cover cannot intersect more than $a_n^{-\delta}$ clusters. So the total number of clusters intersected by this set is no greater than $K_7 \frac{n}{a_n^{-\frac{1}{\eta}+\varepsilon+\delta}}$. Because we can choose ε, δ, we cannot have more than $\frac{n}{a_n^{-\frac{1}{\eta}}}$ clusters that intersect units that satisfy $\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) < a_n^{-\delta}$. Call the number of clusters that intersect this set c_n. Call each cluster $\mathscr{P}_{n,c}$ indexing clusters by c and populations by n.

Consider the following potential outcomes. Let $Y_i(\mathbf{d}) = 0$ when $\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}$. Otherwise, let $Y_i(\mathbf{d}) = y_c d_c$ where $|y_c| < \overline{Y}$ and c is the index of a cluster that unit j belongs to. Then we can write the sum in terms of the clusters:

$$\begin{aligned}
\hat{\theta}_n &= \sum_{i=1}^{n} Y_i(\mathbf{d})\omega_i(\mathbf{d}) \\
&= \sum_{c=1}^{c_n} y_c d_c \left(\sum_{i \in \mathscr{P}_{n,c}} \omega_i(\mathbf{d})\right) \\
&= \sum_{c=1}^{c_n} y_c d_c \omega_c(\mathbf{d})
\end{aligned}$$

Under these potential outcomes:

$$\begin{aligned}
n &= \sum_{i=1}^{n} \mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\} + \sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| \\
\theta_n &= \frac{1}{n}\sum_{i=1}^{n} y_i
\end{aligned}$$

$$= \frac{1}{n} \sum_{c=1}^{C} |\mathscr{P}_{n,c}| y_c$$
$$= \chi_n^{-1} \frac{1}{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}|} \sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| y_c$$

Where χ_n is equal to the sequence below. Notice that χ_n is deterministic, is observed by the researcher, and does not depend on the potential outcomes:

$$\chi_n := \frac{n}{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}|} = \frac{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| + \sum_{i=1}^{n} \mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\}}{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}|}$$

Notice that Lemma 8 guarantees that $\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| = O(n)$ and therefore $\chi_n = O(1)$, a fact that is crucial later.

If we take the hypothesis of the theorem that $\hat{\theta}_n - \theta_n = O_p(a_n)$ and multiply both sides by the sequence χ_n:

$$\hat{\theta}_n - \theta_n = O_p(a_n)$$
$$\sum_{c=1}^{c_n} y_c d_c \omega_c(\mathbf{d}) - \chi_n^{-1} \frac{1}{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}|} \sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| y_c = O_p(a_n)$$
$$\chi_n \sum_{c=1}^{c_n} y_c d_c \omega_c(\mathbf{d}) - \frac{1}{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}|} \sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| y_c = O_p(a_n \chi_n)$$

Now we move χ_n inside the sum and recall that $\chi_n = O(1)$.

$$\sum_{c=1}^{c_n} y_c d_c (\chi_n \omega_c(\mathbf{d})) - \frac{1}{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}|} \sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| y_c = O_p(a_n)$$

Notice that $\frac{1}{\sum_{c=1}^{c_n} |\mathscr{P}_{n,c}|} \sum_{c=1}^{c_n} |\mathscr{P}_{n,c}| y_c$ is a convex combination of c_n deterministic real numbers y_c. Notice that $y_c d_c$ is a Bernoulli(p) sample of the y_c. Finally, notice that $\chi_n \omega_i(\mathbf{d})$ are random weights that do not depend on y_c. It is not possible to construct a set of weights or a covariance matrix for the d_c that is guaranteed to yield an estimator that converges to a specific convex combination of y_c for every set of y_c such that $|y_c| < \overline{Y}$

that converges faster than $O_p(c_n^{-1/2})$. We have already seen that $c_n = O\left(\frac{n}{a_n^{-\frac{1}{\eta}+\varepsilon+\delta}}\right)$. So now we can solve for a bound on a_n.

$$\begin{aligned}
\sqrt{\frac{a_n^{-\frac{1}{\eta}+\varepsilon+\delta}}{n}} &= O(a_n) \\
n &= O\left(a_n^{-2+\frac{1}{\eta}+\varepsilon+\delta}\right) \\
n^{-\frac{1}{2+\frac{1}{\eta}-\varepsilon-\delta}} &= O(a_n)
\end{aligned}$$

Since this holds for arbitrarily small ε, δ, we have that for every $\varepsilon' > 0$,

$$a_n^{-1} n^{-\left(\frac{1}{2+\frac{1}{\eta}}+\varepsilon'\right)} = O(1)$$

Now suppose for the sake of contradiction that for some $\delta' > 0$:

$$n^{\frac{1}{2+\frac{1}{\eta}}+\delta'}(\hat{\theta}_n - \theta_n) = O_p(1) \, \forall \mathbf{Y}$$

Then $n^{\frac{1}{2+\frac{1}{\eta}}+\delta'} = a_n$ and by the previous result we must have for any $\varepsilon' > 0$:

$$\begin{aligned}
n^{\frac{1}{2+\frac{1}{\eta}}+\delta'} n^{-\left(\frac{1}{2+\frac{1}{\eta}}+\varepsilon'\right)} &= O(1) \\
n^{-\varepsilon'+\delta'} &= O(1)
\end{aligned}$$

Which is a contradiction since this cannot be true for all $\varepsilon' > 0$ when $\delta' > 0$ is fixed.

Which gives us that for every $a, b > 0$ there exists $c > 0$ such that:

$$\limsup_{n\to\infty} \sup_{\mathbf{Y}\in\mathscr{Y}_n} \mathbb{P}\left(n^{\frac{1}{2+\frac{1}{\eta}}+a}\left|\hat{\theta}_n - \theta_n\right| > b\right) > c$$

2.11.3 Proof of Theorem 6

First we show the convergence of HT and then we show that the Hajek adjustment does not alter the convergence.

Consider the following sum which we will show is close to the HP estimator:

$$\tilde{\theta}_{n,\kappa_n} \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i(\mathbf{1})T_{i1}}{p^{\phi_n(i,\kappa_n)}} - \frac{Y_i(\mathbf{0})T_{i0}}{(1-p)^{\phi_n(i,\kappa_n)}}\right)$$

Consider the difference between $\tilde{\theta}_{n,\kappa_n}$ and the IPW estimator $\hat{\theta}^{HT}_{n,\kappa_n}$:

$$\tilde{\theta}_{n,\kappa_n} - \hat{\theta}^{HP}_{n,\kappa_n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{(Y_i(\mathbf{1})-Y_i)T_{i1}}{p^{\phi_n(i,\kappa_n)}} - \frac{(Y_i(\mathbf{0})-Y_i)T_{i0}}{(1-p)^{\phi_n(i,\kappa_n)}}\right)$$

Assumption 6 guarantees that:

$$|Y_i(\mathbf{1})-Y_i|T_{i1} \leq K_6\kappa_n^{-\eta}$$
$$|Y_i(\mathbf{0})-Y_i|T_{i0} \leq K_6\kappa_n^{-\eta}$$

We can use this to control the distance between $\tilde{\theta}_{n,\kappa_n}$ and $\hat{\theta}^{HP}_{n,\kappa_n}$. All of these statements hold almost surely.

$$\begin{aligned}
\left|\tilde{\theta}_{n,\kappa_n} - \hat{\theta}^{HP}_{n,\kappa_n}\right| &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{(Y_i(\mathbf{1})-Y_i)T_{i1}}{p^{\phi_n(i,\kappa_n)}} - \frac{(Y_i(\mathbf{0})-Y_i)T_{i0}}{(1-p)^{\phi_n(i,\kappa_n)}}\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i(\mathbf{1})-Y_i|T_{i1}}{p^{\phi_n(i,\kappa_n)}} + \frac{|Y_i(\mathbf{0})-Y_i|T_{i0}}{(1-p)^{\phi_n(i,\kappa_n)}} \\
&\leq \frac{1}{n}\sum_{i=1}^{n}K_6\kappa_n^{-\eta}\left(\frac{1}{p^{\phi_n(i,s)}} + \frac{1}{(1-p)^{\phi_n(i,s)}}\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}K_6\kappa_n^{-\eta}\left(\frac{1}{(p(1-p))^{\phi_n(s)}} + \frac{1}{(p(1-p))^{\phi_n(s)}}\right) \\
&= \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}
\end{aligned}$$

Next we analyze the convergence of $\tilde{\theta}_{n,\kappa_n}$. We will show that it is unbiased for our estimand and that its variance shrinks at a certain rate. Then we will use Chebysev's Inequality to find its rate of convergence to our estimand.

First, notice that the expectation of $\tilde{\theta}_{n,\kappa_n}$ equals our estimand: θ_n.

$$\begin{aligned} E\left[\tilde{\theta}_{n,\kappa_n}\right] &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{Y_i(\mathbf{1})T_{i1}}{p^{\phi_n(i,s)}}\right] - \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{Y_i(\mathbf{0})T_{i0}}{(1-p)^{\phi_n(i,s)}}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{1}) - \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{0}) \\ &= \theta_n \end{aligned}$$

Now for the variance of $\tilde{\theta}_{n,\kappa_n}$. For ease of notation, define $\tilde{Z}_{i,\kappa_n}$ as the summands of $\tilde{\theta}_{n,\kappa_n}$:

$$\begin{aligned} \tilde{Z}_{i,\kappa_n} &:= \left(\frac{Y_i(\mathbf{1})T_{i1}}{p^{\phi_n(i,\kappa_n)}} - \frac{Y_i(\mathbf{0})T_{i0}}{(1-p)^{\phi_n(i,\kappa_n)}}\right) \\ \tilde{\theta}_{n,\kappa_n} &= \frac{1}{n}\sum_{i=1}^{n} \tilde{Z}_{i,\kappa_n} \end{aligned}$$

We can then write the variance as:

$$Var\left(\tilde{\theta}_{n,\kappa_n}\right) = \frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1} \text{Cov}(\tilde{Z}_{i,\kappa_n}, \tilde{Z}_{j,\kappa_n})$$

The variance of a random variable cannot exceed the square of its maximum absolute value. Since Assumption 9 stipulates that potential outcomes are uniformly bounded by $\overline{Y}$ and T_{i1}, T_{i0} are mutually exclusive indicators:

$$\begin{aligned} Var(\tilde{Z}_{i,\kappa_n}) &= Var\left(\frac{Y_i(\mathbf{1})T_{i1}}{p^{\phi_n(s)}} + \frac{Y_i(\mathbf{0})T_{i0}}{(1-p)^{\phi_n(s)}}\right) \\ &\leq \frac{\overline{Y}^2}{\min\left(p^{2\phi_n(\kappa_n)}, (1-p)^{2\phi_n(\kappa_n)},\right)} \end{aligned}$$

$$\leq \frac{\overline{Y}^2}{(p(1-p))^{2\phi_n(s)}}$$

Since covariances cannot exceed the maximum of the two variances:

$$\text{Cov}(\tilde{Z}_{i,\kappa_n}, \tilde{Z}_{j,\kappa_n}) \leq \frac{\overline{Y}^2}{(p(1-p))^{2\phi_n(\kappa_n)}}$$

We can bound the sum of covariances by noticing that if the s-neighorhoods of i, j do not share a cluster, then T_{i1}, T_{i0} is independent of T_{j1}, T_{j0}.

$$\sum_{\mathcal{N}_n(i,\kappa_n)\mathcal{N}_n(j,\kappa_n)\text{don't share a cluster}} \text{Cov}(\tilde{Z}_{i,\kappa_n}, \tilde{Z}_{j,\kappa_n}) = 0$$

How many pairs of neighborhoods touch? We can upper bound this by summing over the clusters. Each cluster can connect no more than the square of the number of neighborhoods intersecting that cluster. So there are at most $\sum_{c=1}^{c_n} \gamma_n(c,\kappa_n)^2$ neighborhoods that share a cluster. This helps us bound the covariance terms.

$$\begin{aligned}
Var(\tilde{\theta}_{n,\kappa_n}) &= \frac{1}{n^2} \sum_{i=1}^{N} \sum_{j=1} \text{Cov}(\tilde{Z}_{i,\kappa_n}, \tilde{Z}_{j,\kappa_n}) \\
&\leq \sum_{N_n(i,s)N_n(j,s)\text{share a cluster}} \frac{\overline{Y}^2}{n^2(p(1-p))^{2\phi_n(i,\kappa_n)}} \\
&\leq \frac{\overline{Y}^2 \sum_{c=1}^{c_n} \gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(\kappa_n)}}
\end{aligned}$$

We will now show the convergence of the HT estimator. Consider any sequence $\delta_n > 0$:

$$\begin{aligned}
P\left(\left|\hat{\theta}_{n,\kappa_n} - \theta_n\right| > \delta_n\right) &= P\left(\left|\hat{\theta}_{n,\kappa_n} - \tilde{\theta}_{n,\kappa_n} + \tilde{\theta}_{n,\kappa_n} - \theta_n\right| > \delta_n\right) \\
&\leq P\left(\left|\hat{\theta}_{n,\kappa_n} - \tilde{\theta}_{n,\kappa_n}\right| + \left|\tilde{\theta}_{n,\kappa_n} - \theta_n\right| > \delta_n\right)
\end{aligned}$$

$$\leq P\left(\frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}} + \left|\tilde{\theta}_{n,\kappa_n} - \theta_n\right| > \delta_n\right)$$
$$= P\left(\left|\tilde{\theta}_{n,\kappa_n} - \theta_n\right| > \delta_n - \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}\right)$$

Now we use Chebyshev's Inequality:

$$P\left(\left|\tilde{\theta}_{n,\kappa_n} - \theta_n\right| > \delta_n - \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}\right) \leq \frac{\frac{\overline{Y}^2\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(s)}}}{\left(\delta_n - \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}\right)^2}$$

Now choose any $\varepsilon > 0$ and set $\delta_n = \frac{1}{\sqrt{\varepsilon}}\sqrt{\frac{\overline{Y}^2\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(s)}}} + \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}$. This gives us an upper bound on the convergence of the IPW estimator (recall that this probability is smaller):

$$P\left(\left|\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n\right| > \frac{1}{\sqrt{\varepsilon}}\sqrt{\frac{\overline{Y}^2\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(s)}}} + \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}\right) \leq \varepsilon$$

Which is equivalent to:

$$P\left(\frac{\left|\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n\right|}{\sqrt{\frac{\overline{Y}^2\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(s)}}} + \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}} > \frac{1}{\sqrt{\varepsilon}} + \left(1 - \frac{1}{\sqrt{\varepsilon}}\right)\frac{\frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}}{\sqrt{\frac{\overline{Y}^2\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(s)}}} + \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}}\right) \leq \varepsilon$$

Since the fraction on the right hand size is less than one, we can increase the right-hand-side by writing:

$$P\left(\frac{\left|\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n\right|}{\sqrt{\frac{\overline{Y}^2\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(s)}}} + \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}} > \frac{1}{\sqrt{\varepsilon}} + 1\right) \leq \varepsilon$$

So we have the rate of convergence:

$$\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n = O_p\left(\sqrt{\frac{\overline{Y}^2 \sum_{c=1}^{c_n} \gamma_n(c,\kappa_n)^2}{n^2(p(1-p))^{2\phi_n(s)}}} + \frac{2K_6\kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}\right)$$

We can remove the constant coefficients without changing the result:

$$\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n = O_p\left(\frac{\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,\kappa_n)^2}}{n} + \kappa_n^{-\eta}}{(p(1-p))^{\phi_n(s)}}\right)$$

By Lemma 10, $\phi_n(\kappa_n)$ is bounded. By Lemma 11, $\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,\kappa_n)^2}}{n} = O\left(\sqrt{\frac{\kappa_n}{n}}\right)$.

$$\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n = O_p\left(\sqrt{\frac{\kappa_n}{n}} + \kappa_n^{-\eta}\right)$$

When $\kappa_n = n^{\frac{1}{2\eta+1}}$,

$$\hat{\theta}^{HT}_{n,\kappa_n} - \theta_n = O_p\left(n^{-\frac{1}{2+\frac{1}{\eta}}}\right)$$

Next we show the same result for the Hajek estimator. First we asymptotically approximate the Hajek estimator using a first-order Taylor expansion of $\frac{1}{n}\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})}$ about its expectation of 1:

$$\begin{aligned}
\sum_{i=1}^{n} \omega_i^{HJ} Y_i - \tilde{\theta} = {} & \frac{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} Y_i}{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})}} - \frac{1}{n}\sum_{i=1}^{n} E\left[Y_i \,|\, T_{i1} = 1\right] \\
& - \frac{\sum_{i=1}^{n} \frac{T_{i0}}{P(T_{i0})} Y_i}{\sum_{i=1}^{n} \frac{T_{i0}}{P(T_{i0})}} + \frac{1}{n}\sum_{i=1}^{n} E\left[Y_i \,|\, T_{i0} = 1\right] \\
\frac{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} Y_i}{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})}} - \frac{1}{n}\sum_{i=1}^{n} E\left[Y_i \,|\, T_{i1} = 1\right] = {} & \left(\frac{1}{n}\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} Y_i - E\left[Y_i \,|\, T_{i1} = 1\right]\right) \\
& - \left(\frac{1}{n}\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} - 1\right) \frac{1}{n}\sum_{i=1}^{n} E\left[Y_i \,|\, T_{i1} = 1\right]
\end{aligned}$$

$$+O_p\left(\left(\frac{1}{n}\sum_{i=1}^{n}\frac{T_{i1}}{P(T_{i1})}-1\right)^2\right)$$

We can use the previous result for Horvitz-Thompson estimators to derive: $\left(\frac{1}{n}\sum_{i=1}^{n}\frac{T_{i1}}{P(T_{i1})}-1\right)=O_p(\sqrt{\frac{\kappa_n}{n}})$. To see this, set $Y_i=d_i$. This yields:

$$\frac{\sum_{i=1}^{n}\frac{T_{i1}}{P(T_{i1})}Y_i}{\sum_{i=1}^{n}\frac{T_{i1}}{P(T_{i1})}}-\frac{1}{n}\sum_{i=1}^{n}E\left[Y_i\,|\,T_{i1}=1\right]=\frac{1}{n}\sum_{i=1}^{n}\frac{T_{i1}}{P(T_{i1})}\left(Y_i-E[Y_i\,|\,T_{i1}=1]\right)+O_p\left(\frac{\kappa_n}{n}\right)$$

So the Hajek estimator is asymptotically equivalent to running the Horvtiz-Thompson estimator, resdiualizing, and then rerunning HT on the residuals.

$$\begin{aligned}\sum_{i=1}^{n}\omega_i^{HJ}Y_i-\frac{1}{n}\sum_{i=1}^{n}\left(E\left[Y_i\,|\,T_{i1}=1\right]-E\left[Y_i\,|\,T_{i0}=1\right]\right)&=\frac{1}{n}\sum_{i=1}^{n}\frac{T_{i1}}{P(T_{i1})}\left(Y_i-\frac{1}{n}\sum_{k=1}^{n}E[Y_k\,|\,T_{k1}=1]\right)\\&-\frac{1}{n}\sum_{i=1}^{n}\frac{T_{i0}}{P(T_{i0})}\left(Y_i-\frac{1}{n}\sum_{k=1}^{n}E[Y_k\,|\,T_{k0}=1]\right)\\&+O_p\left(\frac{\kappa_n}{n}\right)\end{aligned}$$

The sum $\frac{1}{n}\sum_{i=1}^{n}\frac{T_{i0}}{P(T_{i0})}\left(Y_i-\frac{1}{n}\sum_{k=1}^{n}E[Y_k\,|\,T_{k0}=1]\right)$ has expectation zero.

Since $\left(Y_i-\frac{1}{n}\sum_{k=1}^{n}E[Y_k\,|\,T_{k0}=1]\right)$ form a set of potential outcomes that satisfy Assumptions 6 and 9, then the variance of this sum must by the previous arguments be $O\left(\frac{\kappa_n}{n}\right)$. By Assumption 6, $\frac{1}{n}\sum_{i=1}^{n}\left(E\left[Y_i\,|\,T_{i1}=1\right]-E\left[Y_i\,|\,T_{i0}=1\right]\right)=\theta_n+O\left(\kappa_n^{-\eta}\right)$. So we have for the Hajek estimator:

$$\hat{\theta}^{HJ}_{n,\kappa_n}-\theta_n=O_p\left(\sqrt{\frac{\kappa_n}{n}}+\kappa_n^{-\eta}\right)$$

When $\kappa_n=n^{\frac{1}{2\eta+1}}$,

$$\hat{\theta}^{HJ}_{n,\kappa_n}-\theta_n=O_p\left(n^{-\frac{1}{2+\frac{1}{\eta}}}\right)$$

2.11.4 Proof of Example 4

Assumption 7: The neighborhood under the minimum distance is the union of the neighborhoods under the ρ_1, ρ_2: $\mathscr{N}_{n,3}(i,s) = \mathscr{N}_{n,1}(i,s) \cup \mathscr{N}_{n,2}(i,s)$. So any covering under ρ_1 is a covering under ρ_3. So if ρ_1 satisfies Assumption 7, then ρ_3 does as well.

Assumption 10: Since $\mathscr{N}_{n,3}(k,s) \subseteq \mathscr{N}_{n,1}(k,Cs)$, if $i,j \in \mathscr{N}_{n,3}(k,s)$, then $\rho_1(i,j) \leq K_{10}Cs$ and thus $\rho_3(i,j) \leq K_{10}Cs$.

Assumption 11: Using the union bound: $|\mathscr{N}_{n,3}(i,s)| \leq |\mathscr{N}_{n,2}(i,s)| + |\mathscr{N}_{n,1}(i,s)| \lesssim s$. Since $\mathscr{N}_{n,1}(i,s) \subseteq |\mathscr{N}_{n,3}(i,s)|$, we have $s \lesssim |\mathscr{N}_{n,1}(i,s)| \leq |\mathscr{N}_{n,3}(i,s)|$

2.11.5 Proof of Corollary 5′

Notice that in the proofs of Theorem 5 and Lemma 8, all of the examples of potential outcomes satisfy Assumption 12. So, the conclusions of the theorem apply even when we restrict to potential outcomes that satisfy this assumption.

2.11.6 Proof of Theorem 7

Define $v_i = b_i - p$. First notice that $\widehat{Cov}(B\mathbf{v}, B\mathbf{v}) = \widehat{Cov}(B\mathbf{b}, B\mathbf{b})$ and $\widehat{Cov}(\mathbf{Y}, B\mathbf{v}) = \widehat{Cov}(\mathbf{Y}, B\mathbf{v})$. Now we show the convergence of $\widehat{Cov}(B\mathbf{v}, B\mathbf{v})$. By Lemma 12:

$$\mathbb{E}\left[\left(\widehat{Cov}(B\mathbf{v}, B\mathbf{v}) - \frac{p(1-p)tr(B'B)}{n}\right)^2\right] = \mathscr{O}\left(\frac{||B'\mathbf{1}||_2^4}{n^4}\right) + \mathscr{O}\left(\frac{\sum_{c=1}^{C_n}(B'B)_{cc}^2 + tr(B'BB'B)}{n^2}\right)$$

We can analyze each of these. First consider the term $\sum_{c=1}^{c_n}(B'B)_{c,c}^2$. Use the fact that $B_{ic} \in [0,1]$, $\sum_{c=1}^{c_n}(B'B)_{c,c}^2 \leq ||B'\mathbf{1}||_2^2$. By hypothesis, B_{ic} is less than one and always zero when cluster c is not intersected by the κ_n neighborhood of unit i. So $\frac{||B'\mathbf{1}||_2}{n} \leq \frac{\sqrt{\sum_{c=1}^{cn}\gamma_n(c,\kappa_n)^2}}{n}$. By Lemma 11, $\frac{\sqrt{\sum_{c=1}^{cn}\gamma_n(c,\kappa_n)^2}}{n} = \mathscr{O}\left(\frac{\max\{\kappa_n, g_n\}}{\sqrt{g_n}\sqrt{n}}\right)$. Now consider the trace. $tr(B'BB'B) = \sum_{i=1}^n\sum_{j=1}^n\left((BB')_{ij}\right)^2$. By assumption, B_{ic} equals zero if cluster c is not within distance κ_n of unit i. So $(BB')_{ij}$ equals zero if i,j do not have κ_n-neighborhoods that intersect a common cluster. In order for i,j to have κ_n-neighborhoods that intersect

a common cluster under a Scaling Clusters design, by Assumption 10, $\rho(i,j), \rho(j,i) \leq K_{10}^2(g_n + 2\kappa_n)$. By Assumption 11, for each unit i, there can be at most $K_{11}K_{10}^2(g_n + 2\kappa_n)$ units that satisfy this. Since $B_{ij} \in [0,K]$, this implies that: $tr(B(B)'B(B')) \leq K^2 n K_{11} K_{10}^2 (g_n + 2\kappa_n)$. Thus, for $\lim_{n\to\infty} g_n n < 1$ we have:

$$\mathbb{E}\left[\left(\widehat{Cov}(B\mathbf{v}, B\mathbf{v}) - \frac{p(1-p)tr(B'B)}{n}\right)^2\right] = \mathscr{O}\left(\frac{\max\{g_n^2, \kappa_n^2\}}{g_n n}\right)$$

Let P be $n \times C$ matrix where $P_{i,c}$ indicates that unit i is in cluster c. Assumption 12 lets us rewrite Y_i as:

$$Y_i(\mathbf{v}) = \beta_0 + (AP)_i\mathbf{v} + \mu_i, \quad \sum_{i=1}^n \mu_i = 0$$

So we have $\widehat{Cov}(\mathbf{Y}, B\mathbf{v}) = \widehat{Cov}(AP\mathbf{v}, B\mathbf{v}) + \widehat{Cov}(\mu, B\mathbf{v})$

Analyzing $\widehat{Cov}(\mu, B\mathbf{v}) = \frac{\mu' B\mathbf{v}}{n}$. This has expectation zero. We bound the variance using Assumption 9: $\mathbb{E}\left[\left(\frac{\mu' B\mathbf{v}}{n}\right)^2\right] = \frac{p(1-p)\mu' BB'\mu}{n^2} \leq p(1-p)\overline{Y}^2 \frac{||B'\mathbf{1}||_2^2}{n^2} = \mathscr{O}\left(\frac{\max\{g_n^2, \kappa_n^2\}}{n g_n}\right)$. By Lemma 12:

$$\mathbb{E}\left[\left(\widehat{Cov}(AP\mathbf{v}, B\mathbf{v}) - \frac{p(1-p)tr(B'AP)}{n}\right)^2\right] = \mathscr{O}\left(\frac{||B'\mathbf{1}||_2^2 ||(AP)'\mathbf{1}||_2^2}{n^4}\right)$$
$$+ \mathscr{O}\left(\frac{\sum_{c=1}^{C_n}(B'AP)_{cc}^2 + tr(B'APB'AP) + tr(P'A'BB'AP)}{n^2}\right)$$

By Assumption 6, $||(AP)'\mathbf{1}||_2^2 \leq n\overline{Y}$. Bounding: $\sum_{c=1}^n (B'AP)_{c,c}^2$.

$$\sum_{c=1}^n (B'AP)_{c,c}^2 = \sum_{c=1}^n \left(\sum_{i=1}^n B'_{c,i}(AP)_{i,c}\right)^2 \leq \overline{Y}^2 \sum_{c=1}^n \left(\sum_{i=1}^n B_{i,c}\right)^2 = \overline{Y}^2 ||B'\mathbf{1}||_2^2$$

Next we bound the sum of the traces.

$$tr(B'APB'AP) + tr(B'APP'A'B) = \sum_{c=1}^{c_n}\sum_{k=1}^{c_n} \left(B'AP_{c,k}\right)^2 + \sum_{c=1}^{c_n}\sum_{k=1}^{c_n} B'AP_{c,k}B'AP_{k,c}$$

If $\mathbf{x}$ is the vector of values of $B'AP_{c,k}$ and $\mathbf{y}$ is the vector of values of $B'AP_{k,c}$, then $tr(B'APB'AP) + tr(B'APP'A'B) = ||\mathbf{x}||_2^2 + \langle \mathbf{x}, \mathbf{y} \rangle$. By Cauchy-Schwartz, this is less than $||\mathbf{x}||_2^2 + ||\mathbf{x}||_2 ||\mathbf{y}||_2$. But since $||\mathbf{x}||_2 = ||\mathbf{y}||_2$, we have: $tr(B'APB'AP) + tr(B'APP'A'B) \leq \sum_{c=1}^{c_n} \sum_{k=1}^{c_n} \left(B'AP_{c,k}\right)^2$. We can now bound this sum using the fact that by Assumptions 9 and 12, the absolute row sums of AP are no greater than $\overline{Y}$.

$$\sum_{c=1}^{c_n} \sum_{k=1}^{c_n} \left(B'AP_{c,k}\right)^2 \leq \sum_{c=1}^{c_n} \left(\sum_{k=1}^{c_n} |B'AP_{c,k}| \right)^2 \leq \sum_{c=1}^{c_n} \left(\sum_{i=1}^{n} \sum_{k=1}^{c_n} B'_{c,i} |AP_{i,k}| \right)^2 \leq \overline{Y}^2 ||B'\mathbf{1}||_2^2$$

Putting all of these results together:

$$\mathbb{E}\left[\left(\widehat{Cov}(\mathbf{Y}, B\mathbf{v}) - \frac{p(1-p)tr(B'AP)}{n} \right)^2 \right] = \mathscr{O}\left(\frac{||B'\mathbf{1}||_2^2}{n^2} \right) = \mathscr{O}\left(\frac{\max\{\kappa_n^2, g_n^2\}}{n g_n} \right)$$

Next use a Taylor expansion of $\frac{1}{\widehat{Cov}(B\mathbf{v}, B\mathbf{v})}$ about $\frac{1}{\frac{p(1-p)tr(B'B)}{n}}$.

Define $X_n \equiv \frac{\widehat{Cov}(\mathbf{Y}, B\mathbf{v})}{p(1-p)tr(B'B)/n} + \frac{\widehat{Cov}(\mathbf{Y}, B\mathbf{v})}{p(1-p)tr(B'B)/n} \left(\frac{p(1-p)tr(B'B)}{n} - \widehat{Cov}(B\mathbf{v}, B\mathbf{v}) \right)$. By Taylor's Theorem:

$$\frac{\widehat{Cov}(\mathbf{Y}, B\mathbf{v})}{\widehat{Cov}(B\mathbf{v}, B\mathbf{v})} - X_n = \mathscr{O}_p\left(\frac{\max\{g_n^2, \kappa_n^2\}}{g_n n} \frac{n^2}{tr(B'B)^2} \right)$$

$$\mathbb{E}\left[\left(X_n - \frac{tr(B'AP)}{tr(B'B)} \right)^2 \right] = \mathscr{O}\left(\frac{\max\{\kappa_n^2, g_n^2\}}{n g_n} \frac{n^2}{tr(B'B)^2} \right)$$

2.11.7 Proof of Theorem 8

First we check that B^U, B^* satisfy the hypotheses of Theorem 7. By inspection of their definitions, both matrices clearly have elements that are non negative, no greater than unity, and equal to zero when unit j is not in a cluster that is intersected by $\mathscr{N}_n(i, \kappa_n)$. Next we show that $B\mathbf{1} \leq K\mathbf{1}$ for some K. This is true for B^U because the row sums are

always equal to unit. To show that this is true for B^*, use Lemma 10.

$$B_i^*\mathbf{1} \leq \frac{\max_i \phi_n(i,\kappa_n)}{\overline{\phi}_n(\kappa_n)} \leq \frac{\max_i \phi_n(i,\kappa_n)}{\min_i \phi_n(i,\kappa_n)} \leq \frac{\left(\overline{K}_{11}g_n+1\right)\left(\overline{K}_{11}K_{10}^2(\kappa_n+2g_n)+1\right)}{\left(\underline{K}_{11}\kappa_n+1\right)\left(\underline{K}_{11}g_n+1\right)}$$

The ratio $\frac{\overline{K}_{11}g_n+1}{\underline{K}_{11}g_n+1}$ is bounded because $g_n \geq 0$. If $\frac{g_n}{\kappa_n}$ is bounded, then $\frac{\overline{K}_{11}K_{10}^2(\kappa_n+2g_n)+1}{\underline{K}_{11}\kappa_n+1}$ must be bounded as well. So $B_i^*\mathbf{1}$ is bounded. So B^*, B^U satisfy all the conditions of Theorem 7.

Next we compute the estimand using B^*.

$$\begin{aligned}
\frac{tr(B^{*\prime}B^*)}{n} &= \frac{\sum_{i=1}^n \sum_{c=1}^{c_n} (B_{ic}^*)^2}{n} = \frac{\sum_{i=1}^n \phi_n(i,\kappa_n)\overline{\phi}^{-1}}{n} = \overline{\phi}^{-1} \\
\frac{tr((B^*)'AP)}{n} &= \frac{\sum_{i=1}^n \sum_{c=1}^{c_n} B_{ic}^*(AP)_{ic}}{n} \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^n A_{ik}\overline{\phi}^{-1} \sum_{c=1}^{c_n} \mathbb{1}\left\{k \in \mathscr{P}_c\right\} \mathbb{1}\left\{\mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_c \neq \emptyset\right\} \\
\frac{tr((B^*)'AP)}{tr((B^*)'B^*)} &= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^n A_{ik} \left(\sum_{c=1}^{c_n} \mathbb{1}\left\{k \in \mathscr{P}_c\right\} \mathbb{1}\left\{\mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_c \neq \emptyset\right\} \right) \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^n A_{ik} \mathbb{1}\left\{\mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_{c(k)} \neq \emptyset\right\}
\end{aligned}$$

By Lemma 10, $\overline{\phi} = O\left(\frac{\kappa_n}{g_n}\right)$. By Assumption 6, $\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^n A_{ik} \mathbb{1}\left\{\mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_{c(k)} \neq \emptyset\right\} = \theta_n + \mathscr{O}\left(\kappa_n^{-\eta}\right)$. So the estimand when we use B^* is: $\frac{tr((B^*)'AP)}{tr((B^*)'B^*)} = \theta_n + \mathscr{O}\left(\kappa_n^{-\eta}\right)$

Next we compute the estimand when we use B^U. Computing traces:

$$\begin{aligned}
\frac{tr((B^U)'AP)}{n} &= \frac{1}{n}\sum_{i=1}^n \sum_{c=1}^{c_n} \frac{\sum_{j=1}^n \mathbb{1}\left\{j \in \mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_c\right\}}{|\mathscr{N}_n(i,\kappa_n)|} \sum_{k=1}^n A_{ik} \mathbb{1}\left\{k \in \mathscr{P}_c\right\} \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^n A_{ik} \sum_{c=1}^{c_n} \mathbb{1}\left\{k \in \mathscr{P}_c\right\} \frac{\sum_{j=1}^n \mathbb{1}\left\{j \in \mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_c\right\}}{|\mathscr{N}_n(i,\kappa_n)|} \\
\frac{tr((B^U)'B^U)}{n} &= \frac{\sum_{i=1}^n \sum_{c=1}^{c_n} (B_{ci}^U)^2}{n} \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{c=1}^{c_n} \left(\frac{\sum_{j=1}^n \mathbb{1}\left\{j \in \mathscr{N}_n(i,\kappa_n) \cap \mathscr{P}_c\right\}}{|\mathscr{N}_n(i,\kappa_n)|} \right)^2
\end{aligned}$$

So the estimand when OLS is run with B^U is:

$$\frac{tr((B^U)'AP)}{tr((B^U)'B^U)} = \sum_{i=1}^{n}\sum_{k=1}^{n} A_{ik}\left(\frac{\sum_{c=1}^{c_n}\mathbb{1}\{k\in\mathscr{P}_c\}\frac{\sum_{j=1}^{n}\mathbb{1}\{j\in\mathscr{N}_n(i,\kappa_n)\cap\mathscr{P}_c\}}{|\mathscr{N}_n(i,\kappa_n)|}}{\sum_{m=1}^{n}\sum_{c=1}^{c_n}\left(\frac{\sum_{j=1}^{n}\mathbb{1}\{j\in\mathscr{N}_n(m,\kappa_n)\cap\mathscr{P}_c\}}{|\mathscr{N}_n(m,\kappa_n)|}\right)^2}\right)$$

2.11.8 Proof of Theorem 9

If there is an $\varepsilon > 0$ such that $n^{-\varepsilon}\frac{\kappa_n}{g_n} \to \infty$, then $n^{-\varepsilon}\phi_n \to \infty$ because Assumption 11 guarantees that there are at least $\underline{K}_{11}\kappa_n$ members of each $\mathscr{N}_n(i,\kappa_n)$ but at most $K\overline{K}_{11}g_n$ members of each cluster. If $n^{-\varepsilon}\phi_n \to \infty$ then for each unit i, having $P(T_{1i}=1)$ is for large enough n less than p^{n^ε}. By the Bonferroni inequality $P(\max_i T_{1i}) \leq np^{n^\varepsilon} \to 0$. So if $n^{-\varepsilon}\frac{\kappa_n}{g_n} \to \infty$ then the HT and HJ estimators converge in probability to zero and are inconsistent. So in order to use an IPW estimator, the researcher must set κ_n to be proportional to g_n and neither HT nor HJ can converge at a geometric rate faster than the maximum bias which is $\mathscr{O}\left(g_n^{-\eta}\right) = \mathscr{O}_p(n^{-q\eta})$. Moreover the standard deviation of these estimators is of the order of the square root of the number of clusters or $\mathscr{O}\left(\sqrt{n/g_n}\right) = \mathscr{O}\left(n^{-(1-q)/2}\right)$. Since $q < \frac{1}{2\eta+1}$, $n^{-q\eta}$ is slower than $n^{-(1-q)/2}$. So the bias dominates the variance.

If we set $\kappa_n \propto n^{\frac{1+3q}{4+2\eta}}$ then the bias of any of our estimators decays with $\kappa_n^{-\eta}$ or $n^{-\frac{1+3q}{4/\eta+2}}$. To see that OLS converges with $\mathscr{O}_p\left(\frac{1+3q}{2+4/\eta}\right)$. This can be seen by plugging $g_n = n^q$ and $\kappa_n = n^{\frac{1+3q}{4+2\eta}}$ into the result of Theorem 8. To show that this is faster than ηq, solve the inequality $\frac{1+3q}{4+2\eta} \geq \eta q$ for q and see that this is equivalent to $q \leq \frac{1}{2\eta+1}$, which is the hypothesis of this theorem.

2.11.9 Proof of Remark 13

Consider a setting where units are located on a Euclidean plane and ρ is the square root of the Euclidean distance, clusters are a tessellation of squares of length $n^{\frac{1}{2(2\eta+1)}}$. Define $\tilde{Y}_i = (1-d_i)Y_i(\mathbf{0}) + d_iY_i(\mathbf{1})$. By Theorem 7, $\frac{\widehat{Cov}(B\mathbf{v},\tilde{\mathbf{Y}})}{\widehat{Cov}(B\mathbf{v},B\mathbf{v})} - \theta_n = O_p\left(n^{-\frac{1}{2+1/\eta}}\right)$.

Define a sequence α_n where $\alpha_n \in (0,1)$ and $\alpha_n \to 0$. Consider the set A of all units within distance $(1-\alpha_n)n^{\frac{1}{2(2\eta+1)}}$ of the centerpoint of their own treatment cluster. The members of A are at least distance $\alpha_n n^{\frac{1}{2(2\eta+1)}}$ of the nearest unit with a different treatment status than their own. By Assumption 6, for all members of A, we have $\left|Y_i - \tilde{Y}_i\right| \leq K_6 \alpha_n^{-\eta} n^{\frac{1}{2(\eta+1/\eta)}}$. Moreover, by Assumption 11, there is some constant K such that the fraction of the population not in A is at most $K\alpha_n^2$. So $\left|\widehat{Cov}(B\mathbf{v}, \tilde{\mathbf{Y}}) - \widehat{Cov}(B\mathbf{v}, \mathbf{Y})\right| \lesssim \alpha_n^{-\eta} n^{\frac{1}{2(2+1/\eta)}} + \alpha_n^2$. Finding the α_n that maximizes the rate of decay of this sum yields $\alpha_n = n^{-\frac{1}{2(2+1/\eta)(2+\eta)}}$. This guarantees that $\frac{\widehat{Cov}(B\mathbf{v}, \tilde{\mathbf{Y}})}{\widehat{Cov}(B\mathbf{v}, B\mathbf{v})} - \theta_n = O_p\left(n^{-\frac{1}{(2+1/\eta)(2+\eta)}}\right)$, which is slower than if Assumption 12 had held.

2.12 Additional Proofs

2.12.1 Proof of Theorem 10

This proof involves finding examples of DGPs for many cases. Often we use a strange or pathological DGP as an example. This is because the expected absolute deviation of the Hajek estimator is hard to calculate unless we use a specific DGP. We don't choose pathological counterexamples because those are the only counterexamples. We choose them because they are easiest to analyze.

Recall that given $\mathscr{D}$, Λ_{ij} is the adjacency matrix linking units with neighborhoods that share a cluster.

$$\Lambda_{ij} \equiv 1\left\{\exists c \,|\, \mathscr{P}_c \cap \mathscr{N}_n(i,\kappa) \neq \emptyset \text{ and } \mathscr{P}_c \cap \mathscr{N}_n(j,\kappa) \neq \emptyset\right\}$$

First we show that no pair of $\kappa', \mathscr{D}'$ with a Hajek estimator dominates $\kappa, \mathscr{D}$ paired with a Hajek estimator. This requires handling four cases. These four cases are collectively exhaustive. Ruling out cases 1 and 2 rules out any change in the clustering scheme. Ruling out case 3 means not decreasing κ enough to change the estimator. If the clustering scheme does not change and case 4 is ruled out then κ cannot increase

enough to change the estimator. So the four cases are collectively exhaustive.

Case 1: Fewer clusters

Consider the case where there are fewer clusters under $\kappa', \mathscr{D}'$ than $\kappa, \mathscr{D}$. Consider the DGP where every outcome is equal to $Y_i = d_i$ for all $i \in \mathscr{N}$. So the Hajek estimator is equal to the GATE unless $\prod_{i=1}^{n} T_{i1} = 0$ in which case the absolute deviation is equal to one. Since $\mathscr{D}$ is a scaling clusters design with $\kappa = g$ under $\kappa, \mathscr{D}$ there is at least one unit per cluster with neighborhood entirely contained within that cluster. So $Pr\left(\prod_{i=1}^{n} T_{i1} = 0\right) = p^C$ where C is the number of clusters. Since $\mathscr{D}'$ has fewer clusters this probability must be higher so $\kappa', \mathscr{D}'$ has greater risk.

Case 2: A cluster splits

Consider the case where there exist units $i, j \in \mathscr{N}_n$ that were in the same cluster under $\mathscr{D}$ but are in different clusters under $\mathscr{D}'$. Consider the DGP where all outcomes are equal to $Y_k = cd_k|d_i - d_j|$ where $c > 0$ is small enough to satisfy Assumption 6. Then under $\mathscr{D}$ all outcomes are zero almost surely and the risk of the Hajek estimator is zero. But under $\mathscr{D}'$ there is a positive probability that $d_i \neq d_j$. Conditional on this event any Hajek estimator has a positive probability to equal c unless κ' is large enough such that $d_i \neq d_j$ implies that $T_{k1}, T_{k0} = 0$ for all units k. If this is not the case, then risk is larger under $\mathscr{D}', \kappa'$.

Suppose instead that κ' is large enough such that $d_i \neq d_j$ implies that $T_{k1}, T_{k0} = 0$ for all units k. Then every unit's κ'-neighborhood intersects both clusters under $\mathscr{D}'$ that contain i and j. Consider another DGP where $Y_k = d_k$ for all units k. Under $\mathscr{D}, \kappa$, the Hajek estimator equals the GATE except for the event where all clusters are untreated which has probability at most $(1-p)$ on which it has absolute deviation 1. The risk under $\mathscr{D}, \kappa$ is less than $(1-p)c$. But under $\mathscr{D}', \kappa'$, the Hajek estimator will equal zero and have absolute deviation c if either of the two clusters containing i and j are untreated which happens with probability $1-p^2$ which is greater than $1-p$. So risk under $\mathscr{D}', \kappa'$ is

greater than under $\mathscr{D}, \kappa$.

Case 3: An exposure shrinks

Consider the case where there is a pair of units i, j such that under $\kappa, \mathscr{D}$, $\Lambda_{ij} = 1$, but under $\kappa', \mathscr{D}'$, $\Lambda_{ij} = 0$. Consider the set of potential outcomes where all outcomes are zero except Y_i, which equals zero unless $d_i = 1 - d_j = c$ where c is small enough to satisfy Assumption 6. For this DGP, the estimator under $\kappa, \mathscr{D}$ is almost surely equal to the true GATE of zero but under $\kappa', \mathscr{D}'$, the estimator does not equal zero on the event that $d_i \neq d_j$. So under this DGP the risk is strictly higher under $\kappa', \mathscr{D}'$.

Case 4: κ grows

If we rule out all previous cases and the case where the two design-estimator pairs are equivalent for all DGPs, this means that $\mathscr{D} = \mathscr{D}'$ but $\kappa' > \kappa$ and for at least one unit $\phi_i' > \phi_i$. Since $\mathscr{D} = \mathscr{D}'$ but $\kappa' > \kappa$, every ϕ_i (weakly) is larger under κ'. Consider the DGP where all outcomes are zero except for unit j. Choose unit j such that $\phi_j'/\phi_j = \min_i \phi_i'/\phi$. Let Y_j equal zero unless every unit is treated in which case every outcome equals c where c is small enough such that Assumption 6 is satisfied. Then the Hajek estimator equals zero unless every unit is treated in which case it equals $\frac{c}{\sum_{i=1}^n \frac{1}{P(T_{i1})}}$. Since $\mathscr{D} = \mathscr{D}'$, the probability of every unit being treated is the same under $\mathscr{D}$ and $\mathscr{D}'$ so we need only consider the absolute deviation conditional on the event that everyone is treated which is $c\left(\frac{1}{n} - \frac{1}{p^{\phi_j}\sum_{i=1}^n \frac{1}{P(T_{i1})}}\right) = c\left(\frac{1}{n} - \frac{1}{p^{\phi_j}\sum_{i=1}^n p^{-\phi_i}}\right)$. Because of how we have chosen unit j, $p^{\phi_j}\sum_{i=1}^n p^{-\phi_i} \leq p^{\phi_j'}\sum_{i=1}^n p^{-\phi_i'}$. So the absolute deviation is larger under κ'.

No HT estimator dominates HJ

Next we show that no Horvitz-Thompson estimator dominates the Hajek estimator. To do this we consider two subcases. First consider the subcase where the HT estimator is paired with a cluster-randomized design $\mathscr{D}'$ with at least as many clusters as $\mathscr{D}$. Now consider the DGP where every outcome is equal to zero unless all units are treated in

which case every outcome is equal to c where c is small enough to satisfy Assumption 6. Then the absolute deviation Hajek estimator is equal to c unless either every unit is treated in which case its absolute deviation is equal to zero. The Horvitz Thompson estimator also has absolute deviation equal to c if not every unit is treated. If every unit is treated then the Horvitz Thompson estimate is greater than c/p. So the absolute deviation of HT is larger on this event which has at least as likely under $\mathscr{D}'$. So the expected absolute deviation of HT is higher.

Next consider the subcase where the HT estimator is paired with a cluster-randomized design $\mathscr{D}'$ with strictly fewer clusters than $\mathscr{D}$. Now consider the DGP where all outcomes are equal to $Y_i = d_i$. In this case the Hajek estimator equals the GATE unless all outcomes are untreated. So its expected absolute deviation is equal to p^C where C is the number of clusters under $\mathscr{D}$. The Horvitz-Thompson estimator will also equal zero if no unit is treated but since $\mathscr{D}'$ has fewer clusters this is more likely under $\mathscr{D}'$. So no Horvtitz-Thompson estimator paried with a cluster-randomized design dominates the Hajek estimator paired with a scaling clusters design.

No OLS estimator dominates

Consider the DGP where $Y_i = \mathbf{1}\{\tilde{s}_i < \kappa_n\}\kappa_n^{-\eta} K_6 q_i$ for all units i where $|q_i| < 1$. So the GATE is zero. The Hajek estimator will always equal the GATE and has zero risk. The OLS estimator can be nonzero. For example if $q_i = \frac{cx_i}{(\sum x_i^2)^2}$ where x_i is the demeaned OLS regressor and c is chosen to be small enough such that Assumption 6 is satisfied, then the OLS estimator is positive unless every unit has a fully saturated or fully dissaturated neighborhood. If this happens with probability 1 then OLS equals zero always for any DGP and does not dominate Hajek.

2.12.2 Proof of Theorem 11

This proof has four steps. In step 1, I provide an upper bound on the expected absolute deviation of the Horvitz-Thompson estimator. In step 2, I provide a set of potential outcomes $\overline{Y}^*$ that achieve this bound asymptotically. In step 3, I show that under $\overline{Y}^*$, the Horvitz-Thompson and Hajek estimators are asymptotically equivalent. Step 4 completes the proof by putting these pieces together.

Step 1: Upper bound on risk of HT

First decompose the estimate. Define $\tilde{Y}_i \equiv d_i Y_i(\mathbf{1}) + (1-d_i)Y_i(\mathbf{0})$.

$$\sum_{i=1}^{n} \omega_i Y_i = \sum_{i=1}^{n} \omega_i \tilde{Y}_i + \sum_{i=1}^{n} \omega_i \left(Y_i - \tilde{Y}_i\right)$$

By the triangle inequality:

$$\begin{aligned} E\left[\left|\sum_{i=1}^{n} \omega_i Y_i - \theta_n\right|\right] &\leq E\left[\left|\sum_{i=1}^{n} \omega_i \tilde{Y}_i - \theta_n\right|\right] + E\left[\left|\sum_{i=1}^{n} \omega_i \left(Y_i - \tilde{Y}_i\right)\right|\right] \\ &\leq E\left[\left|\sum_{i=1}^{n} \omega_i \tilde{Y}_i - \theta_n\right|\right] + K_6 \sum_{i=1}^{n} E\left[|\omega_i| \tilde{s}_i^{-\eta}\right] \end{aligned}$$

The second term does not depend on potential outcomes. The first term still depends on potential outcomes. By Theorem 12, $n^{\frac{1}{2+1/\eta}}\left(\sum_{i=1}^{n} \omega_i \tilde{Y}_i - \theta_n\right)$ is asymptotically normal with expectation zero . The expected absolute value of a mean-zero normally distributed random variable is equal to $\sqrt{2/\pi}$ times its standard deviation:

$$n^{\frac{1}{2+1/\eta}} E\left[\left|\sum_{i=1}^{n} \omega_i \tilde{Y}_i - \theta_n\right|\right] = n^{\frac{1}{2+1/\eta}} \sqrt{\frac{2}{\pi} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}\left(\omega_i \tilde{Y}_i, \omega_i \tilde{Y}_j\right)} + o(1)$$

For the Horvitz-Thompson weights the covariance is:

$$n\text{Cov}\left(\omega_i^{HT} \tilde{Y}_i, \omega_i^{HT} \tilde{Y}_j\right)$$

$$= E\left[\frac{T_{i1}Y_i(\mathbf{1})}{P(T_{j1})}\frac{T_{j1}Y_j(\mathbf{1})}{P(T_{j1})} + \frac{T_{i0}Y_i(\mathbf{0})}{P(T_{i0})}\frac{T_{j0}Y_j(\mathbf{0})}{P(T_{j0})} - \frac{T_{i1}Y_i(\mathbf{1})}{P(T_{i1})}\frac{T_{j0}Y_j(\mathbf{0})}{P(T_{j0})} - \frac{T_{i0}Y_i(\mathbf{0})}{P(T_{i0})}\frac{T_{j1}Y_j(\mathbf{1})}{P(T_{j1})}\right]$$
$$- \left(Y_i(\mathbf{1}) - Y_i(\mathbf{0})\right)\left(Y_j(\mathbf{1}) - Y_j(\mathbf{0})\right)$$
$$= \frac{Y_i(\mathbf{1})Y_j(\mathbf{1})}{p^{\phi_{ij}}} + \frac{Y_i(\mathbf{0})Y_j(\mathbf{0})}{p^{\phi_{ij}}} - \mathbf{1}\left\{\phi_{ij} = 0\right\}\left(Y_i(\mathbf{1})Y_j(\mathbf{0}) + Y_i(\mathbf{0})Y_j(\mathbf{1})\right)$$
$$- Y_i(\mathbf{1})Y_j(\mathbf{1}) - Y_i(\mathbf{0})Y_j(\mathbf{0}) + Y_i(\mathbf{1})Y_j(\mathbf{0}) + Y_i(\mathbf{0})Y_j(\mathbf{1}))$$
$$= \mathbf{1}\left\{\phi_{ij} > 0\right\}\left(\left(Y_i(\mathbf{1})Y_j(\mathbf{1}) + Y_i(\mathbf{0})Y_j(\mathbf{0})\right)\left(\frac{1}{p^{\phi_{ij}}} - 1\right) + Y_i(\mathbf{1})Y_j(\mathbf{0}) + Y_i(\mathbf{0})Y_j(\mathbf{1})\right)$$

Each covariance is therefore maximized by:

$$n\text{Cov}\left(\omega_i^{HT}\tilde{Y}_i, \omega_i^{HT}\tilde{Y}_j\right) \leq \mathbf{1}\left\{\phi_{ij} > 0\right\}\overline{Y}^2\frac{2}{p^{\phi_{ij}}}$$

And the expected absolute deviation for the HT estimator is upper bounded by:

$$n^{\frac{1}{2+1/\eta}}E\left[\left|\sum_{i=1}^{n}\omega_i^{HT}\tilde{Y}_i - \theta_n\right|\right] \leq n^{\frac{1}{2+1/\eta}}\frac{2\overline{Y}}{n}\sqrt{\frac{1}{\pi}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\phi_{ij} > 0\right\}\frac{1}{p^{\phi_{ij}}}} + o(1)$$

This gives us the asymptotic upper bound on risk:

$$n^{\frac{1}{2+1/\eta}}E\left[\left|\sum_{i=1}^{n}\omega_i^{HT}Y_i - \theta_n\right|\right]$$
$$\leq n^{\frac{1}{2+1/\eta}}\frac{2\overline{Y}}{n}\sqrt{\frac{1}{\pi}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\phi_{ij} > 0\right\}\frac{1}{p^{\phi_{ij}}}} + n^{\frac{1}{2+1/\eta}}\frac{K_6}{n}\sum_{i=1}^{n}E\left[|\omega_i^{HT}|\bar{s}_i^{-\eta}\right] + o(1)$$

Step 2: Outcomes $\mathbf{Y}^*$ achieve the bound

Next we find a set of potential outcomes $\mathbf{Y}^*(\mathbf{d})$ that achieve the upper bound on HT risk asymptotically. Use Assumption 13 to find two mutually exclusive and collectively exhaustive subsets of the population of equal size $\mathscr{N}_1, \mathscr{N}_2$ such that:

$$\frac{1}{n}\sum_{i\in N_1}\mathbf{1}\left\{\mathscr{N}_n(i,s)\cap N_2\right\} + \frac{1}{n}\sum_{i\in N_2}\mathbf{1}\left\{\mathscr{N}_n(i,s)\cap N_1\right\} \leq K_{13}sn^{-\frac{1}{2}}$$

Given any choice of design $\mathscr{D}$ and radii κ_n define $\mathbf{Y}^*(\mathbf{d})$ as:

$$\tilde{Y}_i^* \equiv \overline{Y}\left(2\mathbf{1}\left\{i \in \mathscr{N}_1\right\} - 1\right)$$
$$Y_i^* \equiv \tilde{Y}_i^*(\mathbf{d}) + \text{sign}\left(\sum_{i=1}^n \omega_i \tilde{Y}_i^*\right)(2d_i - 1)K_6 \tilde{s}_i^{-\eta}$$

Now we show that $\mathbf{Y}^*(\mathbf{d})$ achieves the upper bound on the Horvitz-Thompson risk from step 1. First notice that under these outcomes the GATE is zero, i.e. $\theta_n = 0$. Then notice that $\omega_i(d_i - 1) = |\omega_i|$

$$\begin{aligned} E\left[\left|\sum_{i=1}^n \omega_i^{HT} Y_i^* - \theta_n\right|\right] &= E\left[\left|\sum_{i=1}^n \omega_i^{HT} \tilde{Y}_i^* + \text{sign}\left(\sum_{i=1}^n \omega_i \tilde{Y}_i^*\right) K_6 \sum_{i=1}^n \omega_i^{HT}(2d_i - 1)\tilde{s}_i^{-\eta}\right|\right] \\ &= E\left[\left|\sum_{i=1}^n \omega_i^{HT} \tilde{Y}_i^* + \text{sign}\left(\sum_{i=1}^n \omega_i \tilde{Y}_i^*\right) K_6 \sum_{i=1}^n |\omega_i^{HT}| \tilde{s}_i^{-\eta}\right|\right] \end{aligned}$$

Notice that the two terms inside the absolute value are scalars that always have the same sign. So the triangle inequality holds with equality. Then by the linearity of expectations:

$$E\left[\left|\sum_{i=1}^n \omega_i^{HT} Y_i^* - \theta_n\right|\right] = E\left[\left|\sum_{i=1}^n \omega_i^{HT} \tilde{Y}_i^*\right|\right] + K_6 \sum_{i=1}^n E\left[|\omega_i^{HT}| \tilde{s}_i^{-\eta}\right]$$

By Theorem 12:

$$n^{\frac{1}{2+1/\eta}} E\left[\left|\sum_{i=1}^n \omega_i^{HT} \tilde{Y}_i^*\right|\right] = n^{\frac{1}{2+1/\eta}} \sqrt{\frac{2}{\pi} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}\left(\omega_i \tilde{Y}_i^*, \omega_i \tilde{Y}_j^*\right)} + o(1)$$

If $\phi_{ij} = 0$, then $\text{Cov}\left(\omega_i \tilde{Y}_i^*, \omega_i \tilde{Y}_j^*\right) = 0$. If $\phi_{ij} = 1$ and $i, j \in \mathscr{N}_1$ or $i, j \in \mathscr{N}_2$, then $\text{Cov}\left(\omega_i \tilde{Y}_i^*, \omega_i \tilde{Y}_j^*\right) = \frac{1}{p^{\phi_{ij}}}$. So each covariance is equal to its upper bound unless i and j come from different halves. By Assumption 13, this happens for only O pairs of units.

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}\left(\omega_i\tilde{Y}_i^*,\omega_i\tilde{Y}_j^*\right) \geq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\phi_{ij}>0\right\}\frac{1}{p^{\phi_{ij}}} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\phi_{ij}>0\right\}\mathbf{1}\left\{i\in\mathscr{N}_1\cap j\in\mathscr{N}_2\right\} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\phi_{ij}>0\right\}\mathbf{1}\left\{i\in\mathscr{N}_2\cap j\in\mathscr{N}_1\right\}$$

If $\mathbf{1}\left\{\phi_{ij}>0\right\}\mathbf{1}\left\{i\in\mathscr{N}_1\cap j\in\mathscr{N}_2\right\}$ then since this is a Scaling Clusters design and by Assumption 10, $j\in\mathscr{N}_n(i,K_{10}(\kappa_n+\alpha g_n))$. Invoking Assumption 13 we can show that these two overlap terms disappear:

$$\begin{aligned}
&n^{\frac{1/2}{2+1/\eta}}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\phi_{ij}>0\right\}\left(\mathbf{1}\left\{i\in\mathscr{N}_1\cap j\in\mathscr{N}_2\right\}+\mathbf{1}\left\{i\in\mathscr{N}_2\cap j\in\mathscr{N}_1\right\}\right)\\
&\leq K_{13}(\kappa_n+\alpha g_n))n^{-\frac{3}{2}+\frac{1/2}{2+1/\eta}}\\
&\lesssim n^{\frac{1}{2\eta+1}-\frac{3}{2}+\frac{1/2}{2+1/\eta}}\\
&= n^{\frac{-3\eta-1/2+\eta/2}{2\eta+1}}\\
&= o(1)
\end{aligned}$$

Which means that:

$$n^{\frac{1}{2+1/\eta}}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}\left(\omega_i\tilde{Y}_i^*,\omega_i\tilde{Y}_j^*\right)} = n^{\frac{1}{2+1/\eta}}\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{1}\left\{\phi_{ij}>0\right\}\frac{1}{p^{\phi_{ij}}}}+o(1)$$

So the potential outcomes $\mathbf{Y}^*$ achieve the upper bound on Horvitz-Thompson expected absolute deviation asymptotically:

$$n^{\frac{1}{2+1/\eta}}E\left[\left|\sum_{i=1}^{n}\omega_i^{HT}Y_i^*-\theta_n\right|\right]$$

$$= n^{\frac{1}{2+1/\eta}} \frac{2\overline{Y}}{n} \sqrt{\frac{1}{\pi} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}\left\{\phi_{ij} > 0\right\} \frac{1}{p^{\phi_{ij}}}} + n^{\frac{1}{2+1/\eta}} \frac{K_6}{n} \sum_{i=1}^{n} E\left[|\omega_i^{HT}| \tilde{s}_i^{-\eta}\right] + o(1)$$

Step 3: Equivalence of HT and HJ under $\mathbf{Y}^*$

Now we turn to the Hajek estimator and show that it is asymptotically equivalent to Horvitz-Thompson under $\mathbf{Y}^*$. To see this, first we decompose the estimate into two sums.

$$\begin{aligned} n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} Y_i^* &= n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} \tilde{Y}_i^* + n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} \left(Y_i^* - \tilde{Y}_i^*\right) \\ &= n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} \tilde{Y}_i^* + \text{sign}\left(n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HT} \tilde{Y}_i^*\right) K_6 n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} |\omega_i^{HJ}| \tilde{s}_i^{-\eta} \end{aligned}$$

Next we show that $n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} \tilde{Y}_i^*$ and $n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} \tilde{Y}_i^*$ are asymptotically equivalent. To see this, define the random variables X, Z for convenience.

$$\begin{aligned} \sum_{j=1}^{n} \omega_j^{+} \tilde{Y}_j &= \frac{\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left\{Y_i(\mathbf{1}) = \overline{Y}\right\} \frac{T_{i1}}{P(T_{i1})} - \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left\{Y_i(\mathbf{1}) = -\overline{Y}\right\} \frac{T_{i1}}{P(T_{i1})}}{\frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left\{Y_i(\mathbf{1}) = \overline{Y}\right\} \frac{T_{i1}}{P(T_{i1})} + \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\left\{Y_i(\mathbf{1}) = -\overline{Y}\right\} \frac{T_{i1}}{P(T_{i1})}} \\ &= \frac{X - Z}{X + Z} \end{aligned}$$

Take the Taylor expansion about the expectations of X and Z.

$$\begin{aligned} \frac{X-Z}{X+Z} - \frac{E[X]-E[Z]}{E[X]+E[Z]} &= (X - E[X]) \frac{2E[Z]}{(E[X]+E[Z])^2} - (Z - E[Z]) \frac{2E[X]}{(E[X]+E[Z])^2} \\ &\quad + O((X-E[X])^2) + O((Z-E[Z])^2) + O((X-E[X])(Z-E[Z])) \\ &= (X - E[X])2E[Z] - (Z - E[Z])2E[X] \\ &\quad + O((X-E[X])^2) + O((Z-E[Z])^2) + O((X-E[X])(Z-E[Z])) \end{aligned}$$

Under $\mathbf{Y}^*$, $E[X] = E[Z] = \frac{1}{2}$. This means that,

$$\frac{X-Z}{X+Z} = X - \frac{1}{2} - Z + \frac{1}{2} + O\left(\left(X-\frac{1}{2}\right)^2\right) + O\left(\left(Z-\frac{1}{2}\right)^2\right) + O\left(\left(X-\frac{1}{2}\right)\left(Z-\frac{1}{2}\right)\right)$$

Since $X - \frac{1}{2} = O_p\left(n^{\frac{-1}{2+1/\eta}}\right)$,

$$\frac{X-Z}{X+Z} - (X-Z) = O_p\left(n^{\frac{-2}{2+1/\eta}}\right)$$

But $X-Z$ is the Horvitz-Thompson estimator! This means that:

$$\sum_{i=1}^{n} \omega_i^{HJ} \tilde{Y}_i^* = \sum_{i=1}^{n} \omega_i^{HT} \tilde{Y}_i^* + O_p\left(n^{\frac{-2}{2+1/\eta}}\right)$$

Which implies that:

$$n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} \tilde{Y}_i^* = n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HT} \tilde{Y}_i^* + o_p(1)$$

This also means that as long as $n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HT} \tilde{Y}_i^*$ does not converge in probability to zero:

$$\Pr\left(\text{sign}\left(n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} \tilde{Y}_i^*\right) = \text{sign}\left(n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HT} \tilde{Y}_i^*\right)\right) \to 1$$

Which means that:

$$n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HJ} Y_i^* = n^{\frac{1}{2+1/\eta}} \sum_{i=1}^{n} \omega_i^{HT} Y_i^* + o_p(1)$$

Step 4: Concluding the proof

Steps 1 and 2 show that under $\mathbf{Y}^*$ the risk of HT is asymptotically maximized:

$$\sup_{\mathbf{Y}} n^{\frac{1}{2+1/\eta}} \mathscr{R}\left(\hat{\theta}_n^{HT}, \mathscr{D}, \mathbf{Y}\right) - n^{\frac{1}{2+1/\eta}} \mathscr{R}\left(\hat{\theta}_n^{HT}, \mathscr{D}, \mathbf{Y}^*\right) = o(1)$$

By Step 3:

$$n^{\frac{1}{2+1/\eta}} \mathscr{R}\left(\hat{\theta}_n^{HT}, \mathscr{D}, \mathbf{Y}^*\right) = n^{\frac{1}{2+1/\eta}} \mathscr{R}\left(\hat{\theta}_n^{HJ}, \mathscr{D}, \mathbf{Y}^*\right) + o(1)$$

If $\mathbf{Y}^*$ did not maximize risk up to $o(1)$ then there would be some other set of potential outcomes $\mathbf{Y}'$ under which the minimum of HT and HJ risk is higher by a non-vanishing amount. Since HT and HJ risk are equal under $\mathbf{Y}^*$ by step 3, then $\mathbf{Y}'$ would have to raise the risk of both in order to raise the risk of their minimum. But this is impossible since by steps 1 and 2 the risk of HT cannot go any higher.

2.12.3 Proof of Proposition 3

Using Assumption 12

$$\begin{aligned}
Y_i &= \beta_0 + A\mathbf{d} + \mu_i \\
E[Y_i \mid \mathscr{F}_i] &= \beta_0 + \tilde{A}_i\mathbf{d} + p(A_i - \tilde{A}_i)\mathbf{1} + \mu_i \\
\sum_{i=1}^{n} E[Y_i\omega_i] - \theta_n &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{A}_i - A_i)\mathbf{1} \\
E[Y_i \mid \mathscr{F}_i] &= \tilde{A}_i\mathbf{d} + p(A_i - \tilde{A}_i)\mathbf{1}
\end{aligned}$$

Now using Assumption 6′,

$$\sum_{i=1}^{n} \omega_i \left(Y_i - E[Y_i \mid \mathscr{F}_i]\right) = \sum_{i=1}^{n} \omega_i (A_i - \tilde{A}_i)(\mathbf{d} - p\mathbf{1})$$

$$
\begin{aligned}
[B]+[C] &= \sum_{i=1}^{n}(A_i-\tilde{A}_i)\left(\omega_i(\mathbf{d}-p\mathbf{1})+\mathbf{1}/n\right) \\
|[B]+[C]| &\leq K_6\sum_{i=1}^{n}\min\left\{1,\kappa_n^{-\eta}\right\}||\omega_i(\mathbf{d}-p\mathbf{1})+\mathbf{1}/n||_\infty \\
&\leq K_6\sum_{i=1}^{n}\min\left\{1,\kappa_n^{-\eta}\right\}\max\{p,1-p\}\left(|\omega_i|+1/n\right)
\end{aligned}
$$

The final result comes from the triangle inequality.

2.12.4 Proof of Theorem 12

Let $Z_i \equiv \frac{1}{n}\omega_i E[Y_i \,|\, \mathscr{F}_i] - \frac{1}{n}E[\omega_i Y_i]$. The terms $\mathbb{E}[Z_i]$ have expectation zero and dependency graph Λ. The maximum degree of Λ is $\lesssim \kappa_n + g_n$. So the maximum degree of the dependency graph Ψ_n satisfies: $\Psi_n \lesssim \kappa_n + g_n$.

Since the weights and outcomes are uniformly bounded, all of the moments exist. $E[|Z_i|^3] \lesssim \frac{\overline{\phi}_n^3}{n^3}$ and $E[Z_i^4] \lesssim \frac{\overline{\phi}_n^4}{n^4}$. So we can use Lemma B.2 from Leung (2022b) (a.k.a. Theorem 3.6 from Ross (2011)) to bound the Wasserschein distance between a standardized version of the distribution of $[A]$ and the standard normal distribution. Define $\sigma_n^2 \equiv \mathrm{Var}([A])$. By assumption, $\liminf_{n\to\infty}\sigma_n^2\overline{\phi}_n^{-2}\frac{n}{\kappa_n^2} > 0$ and $\frac{g_n}{\kappa_n} = O(1)$.

$$
\begin{aligned}
d\left([A]/\sigma_n, \mathscr{Z}\right) &\leq \overline{\phi}^3\frac{\Psi_n^2}{\sigma_n^3}\sum_{i=1}^{n}E[|Z_i|^3] + \overline{\phi}^2\frac{\sqrt{28}\Psi_n^{3/2}}{\sqrt{\pi}\sigma_n^2}\sqrt{\sum_{i=1}^{n}E[Z_i^4]} \\
&\lesssim \frac{n^{3/2}\kappa_n^2}{n^2\kappa_n^{3/2}} + \frac{\kappa_n^{3/2}n}{n^{3/2}\kappa_n^2} \lesssim \frac{\sqrt{\kappa_n}}{\sqrt{n}} + \frac{1}{\sqrt{n}\sqrt{\kappa_n}} \to 0
\end{aligned}
$$

2.12.5 Proof of Theorem 13

Variance for HT

For the Horvitz-Thompson weights:

$$
\omega_i^{HT}E[Y_i|\mathscr{F}_i] = \frac{T_{i1}}{P(T_{i1})}E[Y_i \,|\, T_{i1}=1] - \frac{T_{i0}}{P(T_{i0})}E[Y_i \,|\, T_{i0}=1]
$$

$$E[\omega_i^{HT} Y_i] = E[Y_i \mid T_{i1} = 1] - E[Y_i \mid T_{i0} = 1]$$

By substitution:

$$\begin{aligned}
&\sum_{i=1}^{n} \omega_i^{HT} E[Y_i \mid \mathscr{F}_i] - E[\omega_i^{HT} Y_i] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_{i1}}{P(T_{i1})} - 1 \right) E[Y_i \mid T_{i1} = 1] - \left(\frac{T_{i0}}{P(T_{i0})} - 1 \right) E[Y_i \mid T_{i0} = 1]
\end{aligned}$$

By substitution:

$$\begin{aligned}
&E\left[\left(\sum_{i=1}^{n} \omega_i^{HT} E[Y_i \mid \mathscr{F}_i] - E[\omega_i^{HT} Y_i \right)^2 \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} E[Y_i \mid T_{i1} = 1] E[Y_j \mid T_{j1} = 1] \left(\frac{P(T_{i1} T_{j1})}{P(T_{i1}) P(T_{j1})} - 1 \right) \\
&+ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} E[Y_i \mid T_{i0} = 1] E[Y_j \mid T_{j0} = 1] \left(\frac{P(T_{i0} T_{j0})}{P(T_{i0}) P(T_{j0})} - 1 \right) \\
&+ \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} E[Y_i \mid T_{i1} = 1] E[Y_j \mid T_{j0} = 1]
\end{aligned}$$

Let C be the $n \times n$ block diagonal adjacency matrix that links two units if and only if they are in the same cluster. The variance can be further decomposed as:

$$\begin{aligned}
\tilde{\theta}_i &= E[Y_i \mid T_{i1} = 1] - E[Y_i \mid T_{i0} = 1] \\
\tilde{\theta} &= \frac{1}{n} \sum_{i=1}^{n} E[Y_i \mid T_{i1} = 1] - E[Y_i \mid T_{i0} = 1]
\end{aligned}$$

Continuing:

$$\begin{aligned}
&E\left[\left(\sum_{i=1}^{n} \omega_i^{HT} E[Y_i \mid \mathscr{F}_i] - E[\omega_i^{HT} Y_i] \right)^2 \right] \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} E[Y_i \mid T_{i1} = 1] E[Y_j \mid T_{j1} = 1] \left(\frac{P(T_{i1} T_{j1})}{P(T_{i1}) P(T_{j1})} \right)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[Y_i\,|\,T_{i0}=1]E[Y_j\,|\,T_{j0}=1]\left(\frac{P(T_{i0}T_{j0})}{P(T_{i0})P(T_{j0})}\right)\\
&+\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}\tilde{\theta}_i\tilde{\theta}_j\\
\text{Var}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i\,|\,\mathscr{F}_i]\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[Y_i\,|\,T_{i1}=1]E[Y_j\,|\,T_{j1}=1]\left(\frac{P(T_{i1}T_{j1})}{P(T_{i1})P(T_{j1})}\right)\\
&+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[Y_i\,|\,T_{i0}=1]E[Y_j\,|\,T_{j0}=1]\left(\frac{P(T_{i0}T_{j0})}{P(T_{i0})P(T_{j0})}\right)\\
&-\tilde{\theta}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}\tilde{\theta}_i-\tilde{\theta}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}(\tilde{\theta}_i-\tilde{\theta})\\
&-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})(\tilde{\theta}_i-\tilde{\theta})(\tilde{\theta}_j-\tilde{\theta})\\
&-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}(\tilde{\theta}_i-\tilde{\theta})(\tilde{\theta}_j-\tilde{\theta})
\end{aligned}
$$

The first four sums can be estimated with the following consistent Horvitz-Thompson estimators:

$$
\begin{aligned}
&E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}Y_iY_j\left(\frac{T_{i1}T_{j1}}{P(T_{i1})P(T_{j1})}+\frac{T_{i0}T_{j0}}{P(T_{i0})P(T_{j0})}\right)\right]\\
&\quad=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[Y_i\,|\,T_{i1}=1]E[Y_j\,|\,T_{j1}=1]\left(\frac{P(T_{i1}T_{j1})}{P(T_{i1})P(T_{j1})}\right)\\
&\quad\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[Y_i\,|\,T_{i0}=1]E[Y_j\,|\,T_{j0}=1]\left(\frac{P(T_{i0}T_{j0})}{P(T_{i0})P(T_{j0})}\right)
\end{aligned}
$$

Using Slutszy's Theorem:

$$
\begin{aligned}
\hat{\theta}_{HT}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}\omega_i^{HT}Y_i &\to_p \tilde{\theta}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}\tilde{\theta}_i\\
\hat{\theta}_{HT}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}(\omega_i^{HT}Y_i-\hat{\theta}_{HT}) &\to_p \tilde{\theta}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}(\tilde{\theta}_i-\tilde{\theta})
\end{aligned}
$$

The last term is unidentified but guaranteed to be negative because it is a quadratic form in a positive semi-definite matrix C. So if we ignore the last term we are guaranteed

a conservative variance estimator. The second-to-last term is cannot be estimated and is not guaranteed to be negative. Instead we will bound it in the following way using Assumption 6′:

$$\mu_i \equiv E[Y_i \,|\, d_i = 1] - E[Y_i \,|\, d_i = 0]$$

$$|\mu_i - \tilde{\theta}_i| \leq K_1 \tilde{s}_j^{-\eta}$$

$$\begin{aligned} -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})(\tilde{\theta}_i - \tilde{\theta})(\tilde{\theta}_j - \tilde{\theta}) \leq & -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})(\mu_i - \tilde{\theta})(\mu_j - \tilde{\theta}) \\ & + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})|\mu_i - \tilde{\theta}|K_1\tilde{s}_j^{-\eta} \\ & + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})K_1^2\tilde{s}_i^{-\eta}\tilde{s}_j^{-\eta} \end{aligned}$$

While the last two terms are not necessarily guaranteed to vanish asymptotically, they tend to be small in simulation. The first term can be estimated:

$$\begin{aligned} & E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})\left(\frac{2d_i - 1}{p}Y_i - \hat{\theta}_{HT}\right)\left(\frac{2d_j - 1}{p}Y_j - \hat{\theta}_{HT}\right)\right] \\ & = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})(\mu_i - \tilde{\theta})(\mu_j - \tilde{\theta}) \\ & E\left[\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})\left|\frac{2d_i - 1}{p}Y_i - \tilde{\theta}\right|K_1\tilde{s}_j^{-\eta} \,|\, \tilde{s}\right] \\ & = \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})|\mu_i - \tilde{\theta}|K_1\tilde{s}_j^{-\eta} \end{aligned}$$

Putting all these terms together we have an estimator for $\text{Var}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i \,|\, \mathscr{F}_i]\right)$ that is guaranteed to be asymptotically conservative in expectation.

$$\widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i \,|\, \mathscr{F}_i]\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}Y_iY_j\left(\frac{T_{i1}T_{j1}}{P(T_{i1})P(T_{j1})} + \frac{T_{i0}T_{j0}}{P(T_{i0})P(T_{j0})}\right)$$

$$
\begin{aligned}
&- \hat{\theta}_{HT} \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} (2\omega_i^{HT} Y_i - \hat{\theta}_{HT}) \\
&+ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij}) K_1^2 \tilde{s}_i^{-\eta} \tilde{s}_j^{-\eta} \\
&- \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij}) \left(\frac{2d_i - 1}{p} Y_i - \hat{\theta}_{HT} \right) \left(\frac{2d_j - 1}{p} Y_j - \hat{\theta}_{HT} \right) \\
&+ \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij}) \left| \frac{2d_i - 1}{p} Y_i - \tilde{\theta}_{HT} \right| K_1 \tilde{s}_j^{-\eta}
\end{aligned}
$$

Next we must show that $n^{\frac{1}{2+1/\eta}}$ times the variance of the estimate of the variance converges to zero. By Slutsky's Theorem and consistency we can ignore variance coming from $\hat{\theta}_{HT}$ Notice that each sum can be written as the following sum for a deterministic $n \times n$ symmetric adjacency matrix A random vectors X, U.

$$
\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} X_i U_j
$$

The variance of this sum is:

$$
Var\left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} X_i X_j \right) = \frac{1}{n^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} A_{ij} A_{k\ell} Cov(X_i U_j, X_k U_\ell)
$$

X_i, U_j must be uniformly bounded. In addition, if $A_{i\ell} = A_{kj} = A_{ik} = A_{j\ell} = 0$ then $Cov(X_i U_j, X_k U_\ell) = 0$. That is, $A_{ij} A_{k\ell} Cov(X_i U_j, X_k U_\ell)$ is nonzero only if i, j are neighbors and j, k are neighbors, and either i or j is a neighbor of j or k.

$$
Var\left(\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_{ij} X_i X_j \right) \lesssim \frac{1}{n^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{\ell=1}^{n} A_{ij} A_{k\ell} (A_{ik} + A_{i\ell} + A_{jk} + A_{j\ell} + A_{k\ell})
$$

In the sums above $A = \Lambda$ or $A = \Lambda - C$. So for each unit we can use Assumption

10:

$$A_{ij}A_{k\ell}A_{ik} = 1 \implies \rho(i,j), \rho(k,\ell)\rho(i,k) \leq K_{10}(Kg_n + \kappa_n)$$
$$\implies \rho(j,k) \leq K_{10}^2(Kg_n + \kappa_n)$$

Applying this multiple times we must have:

$$A_{ij}A_{k\ell}(A_{ik} + A_{i\ell} + A_{jk} + A_{j\ell} + A_{k\ell}) = 1 \implies \rho(i,j), \rho(i,k)\rho(i,\ell) \leq K_{10}^2(Kg_n + \kappa_n)$$
$$\implies j,k,\ell \in \mathscr{N}_n\left(i, K_{10}^2(Kg_n + \kappa_n)\right)$$

By assumption 11,

$$\left|\mathscr{N}_n\left(i, K_{10}^2(Kg_n + \kappa_n)\right)\right| \leq \overline{K}_{11}K_{10}^2(Kg_n + \kappa_n)$$

Which implies:

$$n^{\frac{2}{2+1/\eta}}\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{\ell=1}^{n}A_{ij}A_{k\ell}(A_{ik} + A_{i\ell} + A_{jk} + A_{j\ell} + A_{k\ell}) \lesssim \frac{n^{\frac{3}{2+1/\eta}}}{n^4} \to 0$$

Thus,

$$n^{\frac{2}{2+1/\eta}}Var\left(\widehat{Var}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}X_iX_j\right)\right) \to 0$$

Now combining Chebyshev's Inequality with the fact that the variance estimator is conservative in expectation and the $n^{\frac{1}{2+1/\eta}}$-consistency of the Horvitz Thompson estimator yields:

$$\liminf_{n\to\infty} n^{\frac{1}{2+1/\eta}}\left(\widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i \,|\, \mathscr{F}_i]\right) - \text{Var}\left(\sum_{i=1}^{n}\omega_i^{HT}E[Y_i \,|\, \mathscr{F}_i]\right)\right) \geq 0$$

This argument will work for the variance estimators for Hajek and OLS as well. What remains to show is that those variance estimators are asymptotically conservative.

Variance for Hajek

First we asymptotically approximate the Hajek estimator using a first-order Taylor expansion:

$$\begin{aligned}
\sum_{i=1}^{n} \omega_i^{HJ} Y_i - \tilde{\theta} &= \frac{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} Y_i}{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})}} - \frac{1}{n} \sum_{i=1}^{n} E\left[Y_i \,|\, T_{i1} = 1\right] \\
&\quad - \frac{\sum_{i=1}^{n} \frac{T_{i0}}{P(T_{i0})} Y_i}{\sum_{i=1}^{n} \frac{T_{i0}}{P(T_{i0})}} + \frac{1}{n} \sum_{i=1}^{n} E\left[Y_i \,|\, T_{i0} = 1\right] \\
\frac{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} Y_i}{\sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})}} - \frac{1}{n} \sum_{i=1}^{n} E\left[Y_i \,|\, T_{i1} = 1\right] &= \left(\frac{1}{n} \sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} Y_i - E[Y_i \,|\, T_{i1} = 1] \right) \\
&\quad - \left(\frac{1}{n} \sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} - 1 \right) \frac{1}{n} \sum_{i=1}^{n} E\left[Y_i \,|\, T_{i1} = 1\right] + O_p\left(n^{\frac{-2}{2+1/\eta}}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} \left(Y_i - E[Y_i \,|\, T_{i1} = 1]\right) + O_p\left(n^{\frac{-2}{2+1/\eta}}\right)
\end{aligned}$$

So the Hajek estimator is asymptotically equivalent to running the Horvtiz-Thompson estimator, resdiualizing, and then rerunning HT on the residuals.

$$\begin{aligned}
&\sum_{i=1}^{n} \omega_i^{HJ} Y_i - \frac{1}{n} \sum_{i=1}^{n} \left(E\left[Y_i \,|\, T_{i1} = 1\right] - E\left[Y_i \,|\, T_{i0} = 1\right]\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{T_{i1}}{P(T_{i1})} \left(Y_i - \frac{1}{n} \sum_{k=1}^{n} E[Y_k \,|\, T_{k1} = 1] \right) \\
&- \frac{1}{n} \sum_{i=1}^{n} \frac{T_{i0}}{P(T_{i0})} \left(Y_i - \frac{1}{n} \sum_{k=1}^{n} E[Y_k \,|\, T_{k0} = 1] \right) + O_p\left(n^{\frac{-2}{2+1/\eta}}\right)
\end{aligned}$$

The expression for the variance of Hajek is therefore equal to the expression of the variance of Horvtiz-Thompson with $E[Y_i \,|\, T_{i1} = 1] - \frac{1}{n} \sum_{k=1}^{n} E[Y_k \,|\, T_{k1} = 1]$ substituted

for $E[Y_i \mid T_{i1}=1]$ and $E[Y_i \mid T_{i0}=1] - \frac{1}{n}\sum_{k=1}^{n} E[Y_k \mid T_{k0}=1]$ substituted for $E[Y_i \mid T_{i0}=1]$ and $\tilde{\theta}_i - \tilde{\theta}$ substituted for $\tilde{\theta}_i$ and zero substituted for $\tilde{\theta}$. The result is:

$$\begin{aligned}
\delta_{i1} &\equiv E[Y_i \mid T_{i1}=1] - \frac{1}{n}\sum_{k=1}^{n} E[Y_k \mid T_{k1}=1] \\
\operatorname{Var}\left(\sum_{i=1}^{n} \omega_i^{HJ} E[Y_i \mid \mathscr{F}_i]\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Lambda_{ij}\delta_{i1}\delta_{j1}\left(\frac{P(T_{i1}T_{j1})}{P(T_{i1})P(T_{j1})}\right) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Lambda_{ij}\delta_{i0}\delta_{j0}\left(\frac{P(T_{i0}T_{j0})}{P(T_{i0})P(T_{j0})}\right) \\
&\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij})(\tilde{\theta}_i - \tilde{\theta})(\tilde{\theta}_j - \tilde{\theta}) \\
&\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C_{ij}(\tilde{\theta}_i - \tilde{\theta})(\tilde{\theta}_j - \tilde{\theta}) + O\left(n^{\frac{-3}{2+1/\eta}}\right)
\end{aligned}$$

This variance can be estimated (conservatively) in a similar way to the HT variance:

$$\begin{aligned}
r_{1i} &\equiv Y_i - \frac{1}{n}\sum_{k=1}^{n} \frac{T_{k1}Y_k}{P(T_{k1})} \\
r_{0i} &\equiv Y_i - \frac{1}{n}\sum_{k=1}^{n} \frac{T_{k0}Y_k}{P(T_{k0})} \\
\widehat{\operatorname{Var}}\left(\sum_{i=1}^{n} \omega_i^{HJ} E[Y_i \mid \mathscr{F}_i]\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Lambda_{ij} r_{i1} r_{i0}\left(\frac{T_{i1}T_{i0}}{P(T_{i1})P(T_{j1})}\right) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \Lambda_{ij} r_{0i} r_{0j}\left(\frac{T_{i0}T_{j0}}{P(T_{i0})P(T_{j0})}\right) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij}) K_1^2(\tilde{s}_i^{-\eta})(\tilde{s}_j^{-\eta}) \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij})\left(d_i r_{i1} - (1-d_i) r_{i0}\right)\left(d_j r_{j1} - (1-d_j) r_{j0}\right) \\
&\quad + \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (\Lambda_{ij} - Q_{ij})\left|d_i r_{i1} - (1-d_i) r_{i0}\right| K_1(\tilde{s}_j^{-\eta})
\end{aligned}$$

Variance for OLS

First define the following terms for ease of notation:

$$\begin{aligned} x_i &\equiv \overline{\phi}^{-1} B_i(\mathbf{d} - p\mathbf{1}) \\ &= \overline{\phi}^{-1} B_i \mathbf{v} \\ y_i &\equiv E[Y_i \,|\, \mathscr{F}_i] \\ \sum_{i=1}^{n} \omega_i^{OLS} \tilde{Y}_i &= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \\ \tilde{\theta} &\equiv \frac{\frac{1}{n}\sum_{i=1}^{n} E[x_i y_i]}{\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^2\right]} \end{aligned}$$

Using linearity:

$$\begin{aligned} Y_i &= \beta_0 + A_i \mathbf{v} + \mu_i, \quad \sum_{j=1}^{n} \mu_i = 0 \\ \tilde{A}_{ij} &\equiv \Lambda_{ij} A_{ij} \\ y_i &= \beta_0 + E[A_i \mathbf{v} \,|\, \mathscr{F}_i] + \mu_i \\ &= \beta_0 + x_i \overline{\phi} \frac{\tilde{A}_i \mathbf{1}}{B_i \mathbf{1}} + p(A_i - \tilde{A}_i)\mathbf{1} + \mu_i \end{aligned}$$

Which makes the expectations:

$$\begin{aligned} E[x_i] &= 0 \\ E[x_i^2] &= p(1-p) B_i \mathbf{1} \\ E[x_i y_i] &= E[x_i^2] \overline{\phi}^{-2} \frac{\tilde{A}_i \mathbf{1}}{B_i \mathbf{1}} \\ E[x_i y_i] &= p(1-p) \overline{\phi}^{-1} \tilde{A}_i \mathbf{1} \\ \frac{1}{n}\sum_{i=1}^{n} E[x_i y_i] &= p(1-p) \overline{\phi}^{-1} \frac{1}{n} \tilde{A}_i \mathbf{1} \end{aligned}$$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} E[x_i^2] &= p(1-p)\overline{\phi}^{-2}\frac{1}{n}B_i\mathbf{1} \\
&= p(1-p)\overline{\phi}^{-1} \\
\frac{\frac{1}{n}\sum_{i=1}^{n} E[x_i y_i]}{\frac{1}{n}\sum_{i=1}^{n} E[x_i^2]} &= \frac{1}{n}\tilde{A}_i\mathbf{1} \\
&\equiv \tilde{\theta}
\end{aligned}$$

Similarly to Hajek, we can approximate the OLS estimator using a first-order Taylor expansion:

$$\begin{aligned}
\sum_{i=1}^{n} \omega_i^{OLS} E[Y_i \,|\, \mathscr{F}_i] - \tilde{\theta} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i y_i - E[x_i y_i])}{\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^2\right]} - \frac{\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - E\left[x_i^2\right]\right)}{\left(\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^2\right]\right)^2}\frac{1}{n}\sum_{i=1}^{n} E[x_i y_i] \\
&\quad + O_p\left(n^{\frac{-2}{2+1/\eta}}\right) \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - \tilde{\theta}x_i\right)}{\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^2\right]} + O_p\left(n^{\frac{-2}{2+1/\eta}}\right)
\end{aligned}$$

This means that we can approximate the variance as:

$$\begin{aligned}
r_i &\equiv x_i\left(y_i - x_i\tilde{\theta}\right) \\
E\left[\frac{\frac{1}{n}\sum_{i=1}^{n} x_i\left(y_i - x_i\tilde{\theta}\right)}{\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^2\right]}\right] &= 0 \\
\mathrm{Var}\left(\sum_{i=1}^{n} \omega_i^{OLS}\tilde{Y}_i\right) &= \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[r_i r_j]}{\left(\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^2\right]\right)^2} + O_p\left(n^{\frac{-3}{2+1/\eta}}\right)
\end{aligned}$$

Now use the fact that $\sum_{i=1}^{n} E[r_i] = 0$,

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[r_i r_j] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[r_i]E[r_j] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(E[r_i r_j] - E[r_i]E[r_j]) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[r_i r_j] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[r_i]E[r_j]
\end{aligned}$$

The first sum can be estimated easily but the second sum is not identified.

$$\begin{aligned} E[r_i] &= E\left[x_i E[y_i - x_i\tilde{\theta} \,|\, x_i]\right] \\ &= E\left[x_i^2(\tilde{A}_i\mathbf{1} - \tilde{\theta})\right] \\ &= E\left[x_i^2\right](\tilde{\theta}_i - \tilde{\theta}) \end{aligned}$$

Which makes the problematic term:

$$\begin{aligned} -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[r_i]E[r_j] &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}E[x_i^2]E[x_j^2](\tilde{\theta}_i - \tilde{\theta})(\tilde{\theta}_j - \tilde{\theta}) \\ &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})E[x_i^2]E[x_j^2](\tilde{\theta}_i - \tilde{\theta})(\tilde{\theta}_j - \tilde{\theta}) \\ &\quad -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}E[x_i^2]E[x_j^2](\tilde{\theta}_i - \tilde{\theta})(\tilde{\theta}_j - \tilde{\theta}) \end{aligned}$$

The second sum is guaranteed to be negative because C is positive-semi-definite. We need only construct and estimate a bound on the first sum. This can be done with the same Horvitz-Thompson estimator as in the previous part. This results in the following variance estimator. Keep in mind that $E[x_i^2]$ is always known to the researcher because it is an RCT.

$$\begin{aligned} D &\equiv \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n}E\left[x_i^2\right]\right)^2} \\ \widehat{\text{Var}}\left(\sum_{i=1}^{n}\omega_i^{OLS}E[Y_i \,|\, \mathscr{F}_i]\right) &= \frac{D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Lambda_{ij}r_i r_j \\ &\quad + \frac{D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})K_1^2(\tilde{s}_i^{-\eta} + \kappa_n^{-\eta})(\tilde{s}_j^{-\eta} + \kappa_n^{-\eta}) \\ &\quad - \frac{D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij} - Q_{ij})\left(\frac{2d_i - 1}{p}Y_i - \hat{\theta}_{OLS}\right)\left(\frac{2d_j - 1}{p}Y_j - \hat{\theta}_{OLS}\right) \end{aligned}$$

$$+\frac{2D}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\Lambda_{ij}-Q_{ij})\left|\frac{2d_i-1}{p}Y_i-\tilde{\theta}_{OLS}\right|K_1(\tilde{s_j}^{-\eta}+\kappa_n^{-\eta})$$

2.12.6 Proof of Theorem 14

By the same arguments as for Theorem 7, we have that:

$$\begin{aligned}\frac{\widehat{Cov}(B\mathbf{v},AP\mathbf{v})+\widehat{Cov}(B\mathbf{v},\mu)}{\widehat{Cov}(B\mathbf{v},\hat{A}P\mathbf{v})} &= \frac{\frac{tr(BP'A)}{n}+O_p\left(\frac{||B||_1^2+c_n}{n}\right)}{\frac{tr(BP'\hat{A})}{n}+O_p\left(\frac{||B||_1^2+c_n}{n}\right)}\\ &= \frac{\frac{tr(BP'A)}{n}+O_p\left(\frac{||B||_1^2+c_n}{n}\right)}{\frac{tr(BP'\hat{A})}{n}}\end{aligned}$$

Now substitute in B^* use the uniform boundedness of ϕ.

$$\frac{\frac{tr(B_*P'A)}{n}+O_p\left(\frac{||B_*||_1^2+c_n}{n}\right)}{\frac{tr(B_*P'\hat{A})}{n}} = \frac{\frac{\mathbf{1}'A\mathbf{1}}{n}-\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{n}+O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)}{\frac{\mathbf{1}'\hat{A}\mathbf{1}}{n}-\frac{\mathbf{1}'\hat{\tilde{A}}\mathbf{1}}{n}}$$

So the whole estimator is:

$$\widehat{\theta}_n^{SHR} = \frac{\frac{\mathbf{1}'A\mathbf{1}}{n}-\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{n}+O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)}{1-\frac{\mathbf{1}'\hat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\hat{A}\mathbf{1}}}$$

Rearranging and noticing that by Assumption 6, $\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{n}=O(s^{-\eta})$ and by Assumption 14 part 2 $\frac{\mathbf{1}'\hat{\tilde{A}}\mathbf{1}}{n}=O(\kappa_n^{-\eta})$,

$$\widehat{\theta}_n^{SHR} = \frac{\frac{\mathbf{1}'A\mathbf{1}}{n}+O(\kappa_n^{-\eta})+O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)}{1-\frac{O(\kappa_n^{-\eta})}{\mathbf{1}'\hat{A}\mathbf{1}/n}}$$

Using part 4 of Assumption 14:

$$\frac{|\mathbf{1}'\hat{A}\mathbf{1}|}{n} \geq K_{11}$$
$$1 - \frac{O(\kappa_n^{-\eta})}{\mathbf{1}'\hat{A}\mathbf{1}/n} = 1 + o(1)$$

This lets us simplify the rates:

$$\widehat{\theta}_n^{SHR} = \frac{\frac{\mathbf{1}'A\mathbf{1}}{n} + O(\kappa_n^{-\eta}) + O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)}{1 + o(1)}$$
$$\widehat{\theta}_n^{SHR} - \frac{\mathbf{1}'A\mathbf{1}}{n} = O_p\left(\kappa_n^{-\eta} + \frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)$$
$$\widehat{\theta}_n^{SHR} - \theta_n = O_p\left(\kappa_n^{-\eta} + \frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)$$

By Lemma 10, $\phi_n(\kappa_n)$ is bounded. By Lemma 8, $\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n} = O\left(\sqrt{\frac{\kappa_n}{n}}\right)$. When $\kappa_n = n^{\frac{1}{2\eta+1}}$,

$$\hat{\theta}_{n,\kappa_n}^{OLS} - \theta_n = O_p\left(n^{-\frac{1}{2+\frac{1}{\eta}}}\right)$$

2.12.7 Proof of Theorem 15

We start from a line in the proof of Theorem 14:

$$\widehat{\theta}_n^{SHR} = \frac{\frac{\mathbf{1}'A\mathbf{1}}{n} - \frac{\mathbf{1}'\tilde{A}\mathbf{1}}{n} + O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)}{1 - \frac{\mathbf{1}'\hat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\hat{A}\mathbf{1}}}$$
$$= \frac{\frac{\mathbf{1}'A\mathbf{1}}{n} - \frac{\mathbf{1}'\tilde{A}\mathbf{1}}{n} +}{1 - \frac{\mathbf{1}'\hat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\hat{A}\mathbf{1}}} + O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)$$

$$= \frac{\mathbf{1}'A\mathbf{1}}{n}\left(\frac{1-\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{\mathbf{1}'A\mathbf{1}}}{1-\frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}}\right)+O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)$$
$$= \frac{\mathbf{1}'A\mathbf{1}}{n}+\frac{\mathbf{1}'A\mathbf{1}}{n}\left(\frac{\frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}-\frac{\mathbf{1}'\tilde{A}\mathbf{1}}{\mathbf{1}'A\mathbf{1}}}{1-\frac{\mathbf{1}'\widehat{\tilde{A}}\mathbf{1}}{\mathbf{1}'\widehat{A}\mathbf{1}}}\right)+O_p\left(\frac{\sqrt{\sum_{c=1}^{c_n}\gamma_n(c,\kappa_n)^2}}{n}\right)$$

The last term vanishes when κ_n is constant and n goes to infinity.

2.13 Lemmas

2.13.1 Proof of Lemma 8

Choose any $\delta > 0$ and any $\varepsilon \in \left(0, \frac{1}{\eta}\right)$. Now consider the following potential outcomes:

$$Y_i' = \mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\}\left(\frac{a_n^{1-\eta\varepsilon}}{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right)}\sum_{c \text{ s.t.} \mathscr{P}_c \cap \mathscr{N}_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right)} \mathbf{d}_c - \frac{a_n^{1-\eta\varepsilon}}{2}\right)$$

The associated estimand is:

$$\theta_n' = \frac{1}{n}\sum_{i=1}^{n} Y_i'(\mathbf{1}) - Y_i'(\mathbf{0}) = \frac{a_n^{1-\eta\varepsilon}}{n}\sum_{i=1}^{n}\mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\}$$

The expectations of Y_i' is zero

$$E\left[Y_i'\right] = 0$$

And the variance of the Y_i decays quickly. Since the covariance of cluster treatment statuses cannot be positive, the variance is maximized when cluster treatment statuses

are uncorrelated.

$$\text{Var}\left(Y_i'\right) \leq \frac{a_n^{2(1-\eta\varepsilon)} p(1-p)}{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right)} = O\left(a_n^{2(1-\eta\varepsilon)+\delta}\right)$$

By hypothesis we know that:

$$\sum_{i=1}^{n} \omega_i(\mathbf{d}) Y_i' - \theta_n' = O_p\left(a_n\right)$$

$$\sum_{i=1}^{n} \omega_i(\mathbf{d}) Y_i' - \frac{a_n^{1-\eta\varepsilon}}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\} = O_p(a_n)$$

Dividing by $a_n^{1-\eta\varepsilon}$:

$$a_n^{\eta\varepsilon-1} \sum_{i=1}^{n} \omega_i(\mathbf{d}) Y_i' - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\} = O_p(a_n^{\eta\varepsilon})$$

So the same should be true of the expectations:

$$a_n^{\eta\varepsilon-1} \sum_{i=1}^{n} E\left[\omega_i(\mathbf{d}) Y_i'\right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\} = O(a_n^{\eta\varepsilon})$$

$$= o(1)$$

Let's analyze the order of the first term. Since $E[Y_i'] = 0$, the expectation of the product equals the covariance. The covariance can be no greater than the square root of the product of the variances.

$$E\left[\omega_i(\mathbf{d}) Y_i'\right] = \text{Cov}(\omega_i(\mathbf{d}), Y_i')$$

$$|\text{Cov}(\omega_i(\mathbf{d}), Y_i')| \leq \sqrt{\text{Var}(\omega_i(\mathbf{d}))\text{Var}(Y_i')}$$

$$= \frac{1}{4}\sqrt{\text{Var}(\omega_i(\mathbf{d}))} a_n^{1-\eta\varepsilon+\delta/2}$$

Thus,

$$a_n^{\eta\varepsilon-1}\sum_{i=1}^{n} E\left[\omega_i(\mathbf{d})Y_i'\right] \leq a_n^{\delta/2}\sum_{i=1}^{n}\sqrt{\text{Var}(\omega_i(\mathbf{d}))}$$

By hypothesis, $\sum_{i=1}^{n}\sqrt{\text{Var}(\omega_i(\mathbf{d}))} = O_p(1)$. So since $\delta > 0$, $a_n^{\eta\varepsilon-1}\sum_{i=1}^{n} E\left[\omega_i(\mathbf{d})Y_i'\right] = o(1)$. So we have:

$$o(1) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\} = o(1)$$

Which can only be true if $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{\phi_n\left(i, a_n^{-\frac{1}{\eta}+\varepsilon}\right) \geq a_n^{-\delta}\right\} = o(1)$.

2.13.2 Lemma 9

Lemma 9. *If Assumptions 10 and 11 hold, then there is a universal constant C_9 such that the s-packing number of $\mathscr{N}_n(i,as)$ is less than $C_9 a + 1$ for all $s, a > 0, i \in \mathscr{N}_n, n \in \mathbb{N}$.*

Proof. Let $\{\theta_1, \cdots \theta_M\}$ be a maximal s-packing of $\mathscr{N}_n(i,as)$. By maximality, for each unit $j \in \mathscr{N}_n(i,as)$, there is at least one θ_k such that $\min\{\rho(j,\theta_k), \rho(\theta_k, j)\} \leq s$. By Assumption 10, $\max\{\rho(j,\theta_k), \rho(\theta_k,j)\} \leq K_{10}\min\{\rho(j,\theta_k), \rho(\theta_k,j)\} \leq K_{10}s$. So, $\{\theta_1, \cdots \theta_M\}$ are a $K_{10}s$ covering of $\mathscr{N}_n(i,as)$.

Again by Assumption 10, each $\mathscr{N}_n(\theta_k, K_{10}s) \subseteq \mathscr{N}_n(i, K_{10}\max\{a, K_{10}\}s)$.

Now notice that since it is a packing, $\rho(\theta_k, \theta_m) > s$. By the contrapositive of Assumption 10, $\mathscr{N}_n(\theta_k, K_{10}^{-1}s) \cap \mathscr{N}_n(\theta_k, K_{10}^{-1}s) = \emptyset$.

By Assumption 11, each $\mathscr{N}_n(\theta_k, K_{10}^{-1}s)$ contains at least $\underline{K}_{11}K_{10}^{-1}s + 1$ units. All of these are disjoint and must be contained in $\mathscr{N}_n(i, K_{10}\max\{a, K_{10}\}s)$ which itself has by Assumption 11 at most $\overline{K}_{11}K_{10}\max\{a, K_{10}\}s + 1$ units. Since $a, K_{10} \geq 1$ the packing

number M must be less than:

$$M \leq \frac{\overline{K}_{11} K_{10} \max\{a, K_{10}\} s + 1}{\underline{K}_{11} K_{10}^{-1} s + 1} \leq \frac{\overline{K}_{11} K_{10}^2 \max\{a, K_{10}\}}{\underline{K}_{11}} + 1 \leq \frac{\overline{K}_{11} K_{10}^3 a}{\underline{K}_{11}} + 1$$

□

2.13.3 Lemma 10

Lemma 10. Bounding ϕ

If Assumptions 10 and 11 hold, then for a Scaling Clusters Design with sequence g_n,

$$\max\left\{\frac{\underline{K}_{11} s + 1}{\overline{K}_{11} g_n + 1}, 1\right\} \leq \phi_n(i,s) \leq \frac{\overline{K}_{11} K_{10}^2 (s + 2g_n) + 1}{\underline{K}_{11} g_n + 1} \qquad \forall i, n$$

Proof. $\phi_n(i,s)$ is upper bounded by the number of cluster centers with $K_D g_n$ neighborhoods that intersect the s-neighborhood of i. Applying Assumption 10 twice, $\phi_n(i,s)$ is upper bounded by the number of cluster centers inside $\mathcal{N}_n(i, K_{10}^2(s+g_n))$. The g_n-neighborhoods of all of these cluster centers must be contained within $\mathcal{N}_n(i, K_{10}^2(s+2g_n))$. Assumption 11 guarantees that each of these g_n-neighborhoods must contain at least $\underline{K}_{11} g_n + 1$ units. Since this is a scaling clusters design, these g_n-neighborhoods must be mutually exclusive. So there can be at most $\frac{\overline{K}_{11} K_{10}^2 (s+2g_n)+1}{\underline{K}_{11} g_n + 1}$ of these g_n neighborhoods and $\phi_n(i,s) \leq \frac{\overline{K}_{11} K_{10}^2 (s+2g_n)+1}{\underline{K}_{11} g_n + 1}$.

Now for the the lower bound. By assumption 11, each cluster contains at most $\overline{K}_{11} g_n + 1$ units and $\mathcal{N}_n(i,s)$ contains at least $\underline{K}_{11} s + 1$ units. So $\phi_n(i,s) \geq \frac{\underline{K}_{11} s + 1}{\overline{K}_{11} g_n + 1}$ units. □

2.13.4 Lemma 11

Recall that $\gamma_n(c,s)$ denotes the number of s-neighborhoods that intersect cluster c.

Lemma 11. Bounding Cluster-Neighborhood Intersections

If Assumptions 10, and 11 hold, then under a Scaling Clusters design with scaling sequence g_n :

$$\gamma_n(c,s) \lesssim \max\{s, g_n\}, \qquad \forall n \in \mathbb{N} c \in \{1, \cdots c_n\}$$
$$\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,s)^2}}{n} \lesssim \frac{\max\{s, g_n\}}{\sqrt{g_n}\sqrt{n}}, \qquad \forall n \in \mathbb{N}$$

Furthermore, for a sequence κ_n where $g_n \propto \kappa_n$, then:

$$\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,\kappa_n)^2}}{n} \lesssim \sqrt{\frac{\kappa_n}{n}}, \qquad \forall n \in \mathbb{N}$$

Proof. We first bound each individual $\gamma_n(c,s)$. Under a Scaling Clusters design, each cluster is contained within a $K_D g_n$-neighborhood of some unit. If unit q is contained within $\mathscr{N}_n(i, K_D g_n)$ and $\mathscr{N}_n(j,s)$, then $i \in \mathscr{N}_n(q, K_{10} K_D g_n)$ and $j \in \mathscr{N}_n(q, K_{10} s)$. Invoking Assumption 10 again, $j \in \mathscr{N}_n(i, K_{10}^2 \max\{s, K_D g_n\})$. By Assumption 11 there are at most $\overline{K}_{11} K_{10}^2 \max\{s, K_D g_n\} + 1$ units j that meet this criterion for each i. So $\gamma_n(c,s) \leq \overline{K}_{11} K_{10}^2 \max\{s, K_D g_n\} \lesssim \max\{s, g_n\}$.

Now using this bound and the fact that the scaling clusters design requires that each cluster contain a g_n-neighborhood and using Assumption 11, we can conclude that the number of clusters is at most $c_n \lesssim \frac{n}{g_n}$. This lets us bound:

$$\frac{\sqrt{\sum_{c=1}^{c_n} \gamma_n(c,s)^2}}{n} \lesssim \frac{1}{n}\sqrt{c_n \max\{s, g_n\}^2} \lesssim \frac{1}{n}\sqrt{\frac{n}{g_n}} \max\{s, g_n\} = \frac{\max\{s, g_n\}}{\sqrt{g_n}\sqrt{n}}$$

The final result is obtained by substituting in κ_n.

□

2.13.5 Proof of Lemma 12

Lemma 12. Moments of the empirical covariance

Let $\mathbf{v}$ *be a* $C \times 1$ *vector of mean zero i.i.d. random variables. Denote their second moment as* μ_2 *and their fourth moment as* μ_4*. Let* M, W *be* $n \times C$ *matrices of real numbers. Denote* $Q \equiv M'W$ *and assume that* $tr(Q) = O(n)$*. Then,*

$$\widehat{Cov}(W\mathbf{v}, M\mathbf{v}) = \frac{\mu_2 tr(Q)}{n} + O_p\left(\frac{||W'\mathbf{1}||_2||M'\mathbf{1}||_2}{n^2} + \frac{\sqrt{(\mu_4 - 3\mu_2^2)\sum_{i=1}^{C} Q_{i,i}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q)}}{n}\right)$$

Here is the proof.[14]

First, notice that:

$$\begin{aligned}\widehat{Cov}(W\mathbf{v}, M\mathbf{v}) &= \frac{\mathbf{v}'W'M\mathbf{v}}{n} - \frac{\mathbf{1}'W\mathbf{v}}{n}\frac{\mathbf{1}'M\mathbf{v}}{n} \\ &= \frac{\mathbf{v}'Q\mathbf{v}}{n} - \frac{\mathbf{1}'W\mathbf{v}}{n}\frac{\mathbf{1}'M\mathbf{v}}{n}\end{aligned}$$

Since $E[\mathbf{v}] = \mathbf{0}$,

$$E\left[\frac{\mathbf{1}'W\mathbf{v}}{n}\right] = E\left[\frac{\mathbf{1}'M\mathbf{v}}{n}\right] = 0$$

We can bound the variance as well, using the fact that $\mathbf{v}$ are iid:

$$\begin{aligned}Var(\mathbf{1}'W\mathbf{v}) &= \mu_2\mathbf{1}'WIW'\mathbf{1} \\ &= \mu_2\mathbf{1}'WW'\mathbf{1}\end{aligned}$$

[14] Proof sketch for the following section is from lecture notes found here. But these notes contain errors which I have fixed. https://www.math.utah.edu/ davar/math6010/2011/QuadraticForms.pdf

Which gives us the convergence

$$\begin{aligned}\frac{\mathbf{1}'W\mathbf{v}}{n} &= O_p\left(\frac{\sqrt{\mathbf{1}'WW'\mathbf{1}}}{n}\right)\\ &= O_p\left(\frac{||W'\mathbf{1}||_2}{n}\right)\\ \widehat{Cov}(W\mathbf{v},M\mathbf{v}) &= \frac{\mathbf{v}'Q\mathbf{v}}{n} + O_p\left(\frac{||W'\mathbf{1}||_2||M'\mathbf{1}||_2}{n^2}\right)\end{aligned}$$

Now we need only study the convergence of $\frac{\mathbf{v}'Q\mathbf{v}}{n}$. By Chebyshev's Inequality:

$$\frac{\mathbf{v}'Q\mathbf{v}}{n} - \mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right] = O_p\left(\sqrt{Var\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)}\right)$$

Using the fact that $\mathbf{v}$ are mean zero and i.i.d,

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right] &= \frac{\sum_{j=1}^{C}\sum_{i=1}^{C}Q_{ij}\mathbf{v}_i\mathbf{v}_j}{n}\\ &= \frac{\mu_2\sum_{i=1}^{C}Q_{ii}}{n}\\ &= \frac{\mu_2 tr(Q)}{n}\end{aligned}$$

Now we only need to compute $Var\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)$.

$$\begin{aligned}Var\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right) &= \mathbb{E}\left[\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)^2\right] - \mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right]^2\\ &= \mathbb{E}\left[\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)^2\right] - \frac{\mu_2^2 tr(Q)^2}{n^2}\end{aligned}$$

Evaluating the expectation of the squared quadratic form:

$$\mathbb{E}\left[\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{C}\sum_{j=1}^{C}\sum_{k=1}^{C}\sum_{m=1}^{C}Q_{i,j}Q_{k,m}\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k\mathbf{v}_m\right]$$

Notice that each term of the sum falls into one of the following nine categories:

$$\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k\mathbf{v}_m\right] = \begin{cases} \mu_4 & i=j=k=m \\ \mu_2^2 & i=j\neq k=m \\ \mu_2^2 & i=k\neq j=m \\ \mu_2^2 & i=m\neq j=k \\ 0 & \text{otherwise} \end{cases}$$

Now we can rewrite the sum:

$$\begin{aligned} &\sum_{i=1}^{C}\sum_{j=1}^{C}\sum_{k=1}^{C}\sum_{m=1}^{C} Q_{i,j}Q_{k,m}\mathbb{E}\left[\mathbf{v}_i\mathbf{v}_j\mathbf{v}_k\mathbf{v}_m\right] \\ &= \mu_4\sum_{i=1}^{C}Q_{i,i}^2 + \mu_2^2\sum_{i=1}^{C}\sum_{k\neq i}^{C}Q_{i,i}Q_{k,k} + \mu_2^2\sum_{i=1}^{C}\sum_{j\neq i}^{C}Q_{i,j}^2 + \mu_2^2\sum_{i=1}^{C}\sum_{j\neq i}^{C}Q_{i,j}Q_{j,i} \end{aligned}$$

Now we can simplify the first of the double sums:

$$\begin{aligned} \sum_{i=1}^{C}\sum_{k\neq i}^{C}Q_{i,i}Q_{k,k} &= \sum_{i=1}^{C}Q_{i,i}\sum_{k=1}^{C}Q_{k,k} - \sum_{i}^{C}Q_{i,i}^2 \\ &= tr(Q)^2 - \sum_{i=1}^{C}Q_{i,i}^2 \end{aligned}$$

We can simplify the double sum in this way

$$\begin{aligned} \sum_{i=1}^{C}\sum_{j\neq i}^{C}Q_{i,j}^2 &= \sum_{i=1}^{C}\sum_{j=1}^{C}Q_{i,j}^2 - \sum_{i=1}^{C}Q_{i,i}^2 \\ &= tr(Q'Q) - \sum_{i=1}^{C}Q_{i,i}^2 \end{aligned}$$

We can simplify the third double sum in this way:

$$\begin{aligned}\sum_{i=1}^{C}\sum_{j\neq i}^{C} Q_{i,j}Q_{j,i} &= \sum_{i=1}^{C}\sum_{j=1}^{n} Q_{i,j}Q_{j,i} - \sum_{i=1}^{C} Q_{i,i}^2 \\ &= \sum_{i=1}^{C} (Q)_{i,i}^2 - \sum_{i=1}^{C} Q_{i,i}^2 \\ &= tr(Q^2) - \sum_{i=1}^{C} Q_{i,i}^2\end{aligned}$$

Putting these terms together we have:

$$\mathbb{E}\left[\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)^2\right] = (\mu_4 - 3\mu_2^2)\sum_{i=1}^{C} Q_{i,i}^2 + \mu_2^2\left(tr(Q)^2 + tr(Q^2) + tr(Q'Q)\right)$$

We can get rid of the $tr(Q)^2$ term in the variance by noticing that it is equal to the square of the expectation term

$$\mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right]^2 = \frac{\mu_2^2 tr(Q)^2}{n^2}$$

So the variance is:

$$\begin{aligned}Var\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right] &= \mathbb{E}\left[\left(\frac{\mathbf{v}'Q\mathbf{v}}{n}\right)^2\right] - \mathbb{E}\left[\frac{\mathbf{v}'Q\mathbf{v}}{n}\right]^2 \\ &= \frac{(\mu_4 - 3\mu_2^2)\sum_{i=1}^{C} Q_{i,i}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q)}{n^2}\end{aligned}$$

Putting all the pieces together:

$$\begin{aligned}&\widehat{Cov}(W\mathbf{v}, M\mathbf{v}) - \frac{\mu_2 tr(Q)}{n} \\ &= O_p\left(\sqrt{\frac{(\mu_4 - 3\mu_2^2)\sum_{i=1}^{C} Q_{i,i}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q)}{n^2}}\right) + O_p\left(\frac{||W'\mathbf{1}||_2 ||M'\mathbf{1}||_2}{n^2}\right)\end{aligned}$$

If we know that $\frac{\mu_2 tr(Q)}{n} = O(1)$, then we can rewrite as:

$$\begin{aligned}
&\widehat{Cov}(W\mathbf{v}, M\mathbf{v}) \\
&= \frac{\mu_2 tr(Q)}{n} + O_p\left(\sqrt{\frac{(\mu_4 - 3\mu_2^2)\sum_{i=1}^{C} Q_{i,i}^2 + \mu_2^2 tr(Q^2) + \mu_2^2 tr(Q'Q)}{n^2}}\right) \\
&+ O_p\left(\frac{||W'\mathbf{1}||_2 ||M'\mathbf{1}||_2}{n^2}\right)
\end{aligned}$$

2.14 Gravity Trade Model

We provide here a geographic model of trade within a small regional economy that corresponds to Example 5. The model illustrates how the impact of an exogenous cash transfer from outside the region to a subset of the population diffuses throughout the economy. In equilibrium the model induces an "economic distance function" $d(i, j)$ that describes how the impact of the cash transfer to location j on consumption expenditure in location i tapers off as the locations get farther apart. Economic distance increases with spatial distance due to trade costs, but also depends on the relative populations of locations i and j, as well as household tastes for the services that location i exports. This relationship is closely related to the well-known gravity equation, but it describes the causal effects of a cash transfer instead of trade volume.

The economic distance function $d(i, j)$ is not merely a rotation or dilation of Euclidean space. It is asymmetric and does not generally obey the triangle inequality. Yet it retains enough features of Euclidean spatial distance that it satisfies our Assumptions 1-7 from the main paper.

While economic distance increases with spatial distance *ceterus paribus*, the two distance notions induce spaces that the researcher needs to treat quite differently. Since each location produces its own unique tradable service, as locations are added to the economy, household tastes for variety reduce the volume of trade (and increase economic

distance) between every pair of locations. Therefore, adding locations to a circle of fixed diameter will increase geographic density but decrease economic density.

A researcher who wishes to consistently estimate the Global Average Treatment Effect must sample in such a way as to keep density constant in terms of the relevant distance function. The implication is that order to grow the study population while keeping density constant in *economic* space, it is necessary to sample more densely in terms of *geographic* space. This implies asymptotics of "zooming in", i.e. first studying trade across states, then counties, then census blocks, and so on as n grows. This contrasts with a researcher who wishes to combine our estimators with a purely geographic distance function—who would need to increase their sample size by "zooming out", e.g. by adding more and more census tracts.

2.14.1 Model Setup

Consider a region composed of N locations. Each location j contains m_j households and one firm. There are N locally produced services, one for each location, produced exclusively by that location's firm. The numeraire good cannot be produced locally and each household has an endowment of it.

Utility for Household $i \in \{1,2,\cdots N\}$ is Cobb-Douglas in consumption of service j called $C_{i,j}$ and leisure $\ell_i \in [0,1]$. The imported good is F_i. The utility function is identical for both households. Let $\alpha = \sum_{j=1}^{N} \alpha_j$.

$$U_i\left(\{C_{i,j}\}_{j=1}^{N}, F_i, \ell_i\right) = \left(\Pi_{j=1}^{N} C_{i,j}^{\alpha_j}\right) F_i^{\delta} \ell_i^{1-\alpha-\delta} \tag{2.19}$$

All households are endowed with one unit of time and face identical budget constraints. Each household supplies labor only to its own location. Each household is paid wage w_i which differs by location. Consumption of the numeraire good in location i is F_i and the price of services $i \in \{1,2,\cdots N\}$ are denoted p_i. Trade of services

between locations i and j faces iceberg transportation costs $\tau_{i,j} \in (0,1]$ with $\tau_{j,j} = 1$. Each household i has an endowment of one unit of the numeraire good and also receives an exogenous cash transfer T_i denominated in the numeraire imported good.

$$1 + T_i + w_i = F_i + \sum_{j=1}^{N} \left(\frac{p_j}{\tau_{i,j}} \right) C_{ij} + w_i \ell_i \tag{2.20}$$

The N firms indexed $j \in \{1, 2, \cdots N\}$ each produce their own service with production functions equal to labor $1 - \ell_i$. The locally produced services cannot be exported outside the region. Let there be m_i identical households in each location i. The market clearing condition for the service produced in location j is:

$$\sum_{i=1}^{N} m_i C_{i,j} = m_j \left(1 - \ell_j\right) \tag{2.21}$$

Since production is equal to labor, zero profit means that the wage is equal to the price of the service:

$$p_i = w_i$$

Define household expenditure E_i as the sum of consumption of the numeraire good plus spending on services:

$$E_i = F_i + \sum_{j=1}^{N} \left(\frac{p_j}{\tau_{i,j}} \right) C_{ij}$$

To guarantee that an equilibrium exists for all values of the transfer $\mathbf{T}$, we must make the following Assumption 15. The assumption essentially states that no one region is too large or takes up too large a share of the economy's expenditures and that transport costs face their own kind of "triangle inequality."

Assumption 15. *There are constants $M_1, M_3 \in (0,1)$, and $M_2 > 1$ such that for all j:*

- $\frac{1}{n}\sum_{i=1}^{n}\tau_{ij} \leq M_1$ *for all* j
- $\frac{m_j}{m_i} \leq M_2$ *for all* i, j
- $\frac{M_3}{n} \leq \frac{\alpha_j}{\alpha+\delta} \leq \frac{1}{nM_1M_2}$ *For all* j
- $0 < \tau_{ik}\tau_{kj} \leq \tau_{ij}$ *for all* i, j, k
- $\tau_{ii} = 1$ *for all* i

Assumption 16 says that locations exist in two-dimensional Euclidean space and face transport costs that increase with Euclidean spatial distance $\rho(i, j)$. We will see later that the "economic distance" induced by these transportation costs is not spatial (does not satisfy the triangle inequality).

Assumption 16. *Units are located in two-dimensional Euclidean space with Euclidean distance* $\rho(i, j)$ *and* $\tau_{ij} \leq \frac{1}{\rho(i,j)^\gamma}$ *with* $\gamma > 2$.

In order for density in terms of economic distance to remain constant, the geographic density must increase. Assumption 17 requires that the spatial setting is becoming rapidly denser, perhaps through a combination of a smaller spatial setting, denser sampling, and disaggregation of units (smaller m_i). In other words, the researcher is "zooming in" on a smaller and smaller region that they see at higher and higher resolution.

Assumption 17. *There are constants* $M_4 \geq M_5 > 0$ *and* $M_6 > 0$ *such that every circle of radius* s *has at most* $M_4 n^{2\gamma} s^2$ *and at least* $M_5 n^{2\gamma} s^2$ *units inside and for every subset* Q *of the population to construct an* s*-cover in squared Euclidean spatial distance* $\rho(\cdot,\cdot)^2$ *consisting of at most* $M_6 \frac{n^{1-\gamma}}{s}$ *units.*

Theorem 16 says that under Assumption 15 (and no others) there is a sensible equilibrium and that treatment effects are linear in the vector of cash transfers. It also

provides upper and lower bounds on the treatment effects. Direct effects are always larger than unity, implying that the nominal multiplier effect is greater than one. All effects–direct and spillover–must be non-negative. Finally, all effects are bounded above by a quantity that declines as transport costs tighten.

Theorem 16. Linearity of Treatment Effect on Household Expenditure

If Assumption 15 holds, then there exists an $N \times N$ exogenous matrix G with non-negative entries that do not depend on the $N \times 1$ vector of transfers $\mathbf{T}$ such that for every equilibrium, the $N \times 1$ vector of household consumption expenditures $\mathbf{E}$ satisfies:

$$\mathbf{E} = G(\mathbf{T} + \mathbf{1})$$

The elements of G satisfy the bounds:

$$\frac{1}{1 - \frac{\alpha_i}{\alpha+\delta}} \mathbf{1}\{i = j\} \leq G_{ij} \leq \mathbf{1}\{i = j\} + \frac{\alpha_i}{\delta} \frac{m_j}{m_i} \tau_{ij}$$

While the nominal multiplier effect is positive, it is denominated in units of the imported good. Since the relative prices of all the services must rise in response to increasing the transfers, it is not clear whether consumption quantities are all positively affected.

Theorem 17 shows that our model admits a gravity equation similar to those in the trade literature (CITE). This gravity equation is not directly relevant to the problem of estimating the GATE, but is provided here to make explicit the fact that our model belongs to the "gravity" family of trade models.

Theorem 17. Gravity

Let Assumption 15 hold. If X_{ij} is the total expenditure by location i on the service

produced in location j, then for large enough N:

$$X_{ij} = \tau_{ij} \frac{Y_i Y_j}{\sum_{k=1}^{n} Y_k}$$

Theorem 18 defines the "economic distance" function and verifies that spillover effects satisfy Assumption 1 from our main paper with respect to this notion of distance. Notice $d(i,j)$ has several unusual properties. First, since it depends on $\frac{m_i}{m_j}$, it need not satisfy the triangle inequality or be symmetric. This is because a cash transfer to everyone in a big city will affect the small nearby village a great deal, but not vice versa.

Since the taste parameters α_i must sum to $\alpha < 1$, then as n grows, all α_i must shrink. This means that, holding geographic positions and m_i constant, adding more locations to the economy forces $d(i,j)$ to eventually rise for every pair of units because the α_i are shrinking. The purpose of Assumption 17 is to counteract this by requiring the economy to become more geographically dense. This results in an asymptotic framework suited for studying very geographically dense and economically integrated settings. A researcher who did not have an economic model and wished to use spatial distance as their distance notion could not study a sequence of increasing geographically dense populations at all.

Theorem 18. Decay of Spillover Effects

Let $d(i,j) = \left(\frac{\delta}{\alpha_i}\frac{m_i}{m_j}\right)^{2\gamma} \rho(i,j)^2$. If Assumptions 15, 16, and 17 above hold, then $d(\cdot,\cdot)$ satisfies Assumption 1 with $\eta = \gamma/2 - 1$ and $K_1 = M_3 \left(\frac{1}{\delta M_1}\right)^{\gamma/2}$. Neighborhoods induced by $d(\cdot,\cdot)$ also satisfy Assumption 11 with $\underline{K}_{11} = M_5 \left(\frac{1}{\delta M_1}\right)^{\gamma/2}$ and $\overline{K}_{11} = M_4 \left(\frac{1}{\delta M_1}\right)^{\gamma/2}$ and Assumption 10 with $K_{10} = 2\left(\frac{\delta}{(\alpha+\delta)M_3} M_2\right)^{2\gamma} \left(\frac{M_2}{M_3}\right)^{2\gamma}$ and Assumption 7 with $K_7 = M_6 \left(M_2 \frac{\delta}{\alpha+\delta}\right)^{2\gamma}$.

2.14.2 Proofs of Model Results

Proof of Theorem 16

The Cobb-Douglas utility function guarantees that budget shares are spent as:

$$\frac{w_j}{\tau_{i,j}} C_{i,j} = \alpha_j (1 + T_i + w_i)$$

$$w_i \ell_i = (1 - \alpha - \delta)(1 + T_i + w_i)$$

Rewriting consumption:

$$C_{i,j} = \alpha_j \frac{\tau_{i,j}}{w_j} (1 + T_i + w_i)$$

Substituting into the consumption market clearing condition:

$$\sum_i \alpha_j m_i \frac{\tau_{i,j}}{w_j} (1 + T_i + w_i) = m_j (1 - \ell_j)$$

$$\alpha_j \sum_i m_i \tau_{i,j} (1 + T_i + w_i) = m_j w_j - m_j w_j \ell_j$$

$$\alpha_j \sum_i m_i \tau_{i,j} (1 + T_i + w_i) = m_j \left(w_j - (1 - \alpha - \delta)(1 + T_j + w_j) \right)$$

Some more rearranging yields:

$$\frac{m_j}{\alpha_j} (\alpha + \delta) w_j = \sum_{i=1} m_i \tau_{i,j} (1 + T_i) + \frac{m_j}{\alpha_j} (1 - \alpha - \delta)(1 + T_j) + \sum_i m_i \tau_{i,j} w_i$$

We wish to rewrite this into matrix form. Let the $N \times N$ matrix A be diagonal with entries:

$$A_{ii} = \frac{m_i}{\alpha_i}$$

Let the $N \times N$ matrix $\mathscr{T}$ be:

$$\mathscr{T}_{ij} = m_j \tau_{i,j}$$

Now we can rewrite the wage equation in matrix form:

$$\begin{aligned}
(\alpha+\delta)A\mathbf{w} - \mathscr{T}\mathbf{w} &= (\mathscr{T} + (1-\alpha-\delta)A)(\mathbf{1}+\mathbf{T}) \\
((\alpha+\delta)A - \mathscr{T})\,\mathbf{w} &= -((\alpha+\delta)A - \mathscr{T})\,(\mathbf{1}+T) + A(\mathbf{1}+\mathbf{T}) \\
\mathbf{w} &= (((\alpha+\delta)A - \mathscr{T})^{-1}A - I)(\mathbf{1}+T) \\
&= \left(\frac{1}{\alpha+\delta}\left(I - \mathscr{T}A^{-1}\frac{1}{\alpha+\delta}\right)^{-1} - I\right)(\mathbf{1}+\mathbf{T}) \\
&= \left(((\alpha+\delta)A - \mathscr{T})^{-1}A - I\right)(\mathbf{1}+\mathbf{T}) \\
&= D(\mathbf{1}+\mathbf{T})
\end{aligned}$$

We wish to show that this equilibrium is sensible, i.e. that each entry of $\mathbf{w}$ is non-negative. By inspection, each entry of A and $\mathscr{T}$ are non-negative. To show that each entry of $((\alpha+\delta)A - \mathscr{T})^{-1}$ is non-negative it is sufficient to show that $(\alpha+\delta)A - \mathscr{T}$ is diagonally dominant. Diagonal dominance is equivalent to:

$$m_j \frac{\alpha+\delta}{\alpha_j} > \sum_{i=1}^{n} m_i \tau_{i,j}, \qquad \forall_j$$

Now we use Assumption 15. Since the m_j/m_i are bounded above and the limit of sum of τ is also bounded:

$$\sum_{i=1}^{n} \frac{m_i}{m_j}\tau_{i,j} \leq nM_1M_2$$

The second part of Assumption 15 guarantees that:

$$\frac{\alpha_j}{\alpha+\delta} \leq \frac{1}{nM_1M_2}$$

So we have diagonal dominance:

$$m_j \frac{\alpha+\delta}{\alpha_j} > \sum_{i=1}^{n} m_i \tau_{i,j}, \qquad \forall_j$$

Diagonal dominance means that each entry of $((\alpha+\delta)A - \mathscr{T})^{-1}$ is non-negative. This in turn implies that each non-diagonal element of D is also non-negative.

We show that the diagonal elements of $\left(I - \mathscr{T}A^{-1}\frac{1}{\alpha+\delta}\right)^{-1}$ are all greater than $\alpha+\delta$. If $B = \mathscr{T}A^{-1}\frac{1}{\alpha+\delta}$, then this is the same as showing that $I + B_{ii} + (B^2)_{ii} + \cdots$ is greater than $\alpha+\delta$ which must be true since the elements of B are non-negative.

To calculate the effect on expenditure:

$$\begin{aligned}
E_i &= (1 + T_i + w_i) - w_i \ell_i \\
&= (\alpha+\delta)(1 + T_i + w_i) \\
\mathbf{E} &= (\alpha+\delta)(D+\mathbf{I})(\mathbf{1}+\mathbf{T}) \\
\mathbf{E} &= G(\mathbf{1}+\mathbf{T}) \\
G &= (\alpha+\delta)(D+I) \\
&= \left(I - \mathscr{T}A^{-1}\frac{1}{\alpha+\delta}\right)^{-1}
\end{aligned}$$

Let $B = \mathscr{T}A^{-1}\frac{1}{\alpha+\delta}$ or $B_{ij} = \frac{\alpha_i}{\alpha+\delta}\frac{m_j}{m_i}\tau_{ij}$. To lower bound G_{ii} we can use the fact that $G_{ii} = I + B_{ii} + ((B^2)_{ii} + \cdots$. Since the elements of B are non-negative, $G_{ii} \geq 1 + B_{ii} + (B_{ii})^2 + \cdots = \frac{1}{1-B_{ii}} = \frac{1}{1-\frac{\alpha_i}{\alpha+\delta}} > 1$

Now we wish to compute an upper bound. We want to show that: $(B^2)_{ij} \leq cB_{ij}$ where $c < 1$. If this is true, then:

$$\begin{aligned}
(I-B)^{-1}_{ij} &\leq \mathbf{1}\{i=j\} + B_{ij}\frac{1}{1-c} \\
&\leq \mathbf{1}\{i=j\} + \frac{\alpha_i}{\alpha+\delta}\frac{m_j}{m_i}\frac{\tau_{ij}}{1-c}
\end{aligned}$$

Using $\tau_{ik}\tau_{kj} \leq \tau_{ij}$ and $\sum \alpha_k = \alpha$:

$$\begin{aligned}(B^2)_{ij} &= \sum_{k=1}^{n} B_{ik}B_{kj} \\ &\leq \frac{\alpha_i}{\alpha+\delta}\frac{m_j}{m_i}\sum_{k=1}^{n}\frac{\alpha_k}{\alpha+\delta}\tau_{ij} \\ &= \frac{\alpha_i}{\alpha+\delta}\frac{m_j}{m_i}\frac{\alpha}{\alpha+\delta}\tau_{ij} \\ &= \left(\frac{\alpha}{\alpha+\delta}\right)B_{ij}\end{aligned}$$

So we have:

$$\begin{aligned}(I-B)^{-1}_{ij} &\leq \mathbf{1}\{i=j\} + \frac{B_{ij}}{1-\frac{\alpha}{\alpha+\delta}} \\ &= \mathbf{1}\{i=j\} + \frac{\alpha_i}{\left(1-\frac{\alpha}{\alpha+\delta}\right)(\alpha+\delta)}\frac{m_j}{m_i}\tau_{ij} \\ &= \mathbf{1}\{i=j\} + \frac{\alpha_i}{\delta}\frac{m_j}{m_i}\tau_{ij}\end{aligned}$$

This gives us the final bound:

$$\frac{1}{1-\frac{\alpha_i}{\alpha+\delta}}\mathbf{1}\{i=j\} \leq G_{ij} \leq \mathbf{1}\{i=j\} + \frac{\alpha_i}{\delta}\frac{m_j}{m_i}\tau_{ij}$$

Proof of Theorem 17

Total export of service j to location i is:

$$m_i w_j C_{i,j} = m_i \tau_{ij} \alpha_j (1 + T_i + w_i)$$

Define GDP as the amount produced prior to transportation costs:

$$Y_i = m_i(1-\ell_i)w_i$$

We can use Cobb-Douglas budget share to express GDP as the budget share:

$$Y_i = m_i(\alpha+\delta)(1+T_i+w_i)$$

Substituting this in:

$$m_i w_j C_{i,j} = \tau_{ij}\alpha_j \frac{Y_i}{\alpha+\delta}$$

Goal: relate α_j to the size of economy j.

For trade to balance:

$$\begin{aligned} Y_j &= \sum_{k=1}^{N} m_k \frac{w_j}{\tau_{kj}} C_{kj} \\ &= \alpha_j \sum_{k=1}^{N} m_k(1+T_k+w_k) \\ \alpha_j &= (\alpha+\delta)\frac{Y_j}{\sum_{k=1}^{n} Y_k} \end{aligned}$$

Substituting this in:

$$m_i w_j C_{i,j} = \tau_{ij} \frac{Y_i Y_j}{\sum_{k=1}^{n} Y_k}$$

Which is the traditional gravity Equation. We call $X_{ij} = m_i w_j C_{ij}$ the total expenditure on service from location j by location i.

Proof of Theorem 18

We can bound the sum by instead summing over rings and forcing every unit with each ring to be as close to the center as possible:

$$\sum_{j\,:\,d(i,j)>t} G_{ij} \leq \sum_{s=t}^{\infty} |\mathscr{N}_n(i,s+1)\setminus\mathscr{N}_n(i,s)| \max_{j\,:\,d(i,j)\geq s} G_{ij}$$

First we bound the number of units that are not within distance s of unit i but are within d-distance $s-1$. In other words, what the the largest number of units that can be in $\mathscr{N}(i,s) \setminus \mathscr{N}(i,s-1)$.

If $j \in \mathscr{N}(i,s)$, then

$$\left(\frac{\delta}{\alpha_i}\frac{m_i}{m_j}\right)^{2\gamma} \rho(i,j)^2 \leq s$$
$$\rho(i,j) \leq \sqrt{s}\left(\frac{\alpha_i}{\delta}\frac{m_j}{m_i}\right)^{\gamma}$$

A necessary requirement for $j \in \mathscr{N}_n(i,s)$ is:

$$\rho(i,j) \leq \sqrt{s}\left(\frac{1}{n\delta M_1}\right)^{\gamma}$$

The largest number of j that could satisfy this is equal to a constant times $sn^{2\gamma-2\gamma}$. This yields a bound on the neighborhood size:

$$M_5\left(\frac{1}{\delta M_1}\right)^{\gamma/2} s \leq |\mathscr{N}_n(i,s)| \leq M_4\left(\frac{1}{\delta M_1}\right)^{\gamma/2} s$$

This verifies Assumption 11.

We can use our bound on the elements of the matrix G to see that:

$$\begin{aligned}
\max_{j:d(i,j)\geq s} G_{ij} &\leq \max_{j:\left(\frac{\alpha_i}{\delta}\frac{m_j}{m_i}\frac{1}{\rho(i,j)^{\gamma}}\right)^{-2/\gamma} \geq s} \frac{\alpha_i}{\delta}\frac{m_j}{m_i}\frac{1}{\rho(i,j)^{\gamma}} \\
&= \max_{j:\left(\frac{\alpha_i}{\delta}\frac{m_j}{m_i}\frac{1}{\rho(i,j)^{\gamma}}\right) \leq s^{-\gamma/2}} \frac{\alpha_i}{\delta}\frac{m_j}{m_i}\frac{1}{\rho(i,j)^{\gamma}} \\
&= s^{-\gamma/2}
\end{aligned}$$

Using our bounds on the elementso of G and the size of the neighborhoods:

$$\sum_{j:d(i,j)>t} G_{ij} \leq \sum_{s=t}^{\infty} |\mathscr{N}_n(i,s+1) \setminus \mathscr{N}_n(i,s)| \max_{j:d(i,j)\geq s} G_{ij}$$

$$= \sum_{s=t}^{\infty} |\mathscr{N}_n(i,s+1) \setminus \mathscr{N}_n(i,s)| \, s^{-\gamma/2}$$
$$\leq M_3 \left(\frac{1}{\delta M_1}\right)^{\gamma/2} \sum_{s=t}^{\infty} s^{-\gamma/2}$$

Using an integral bound:

$$\sum_{s=t}^{\infty} s^{-\gamma/2} \leq \int_t^{\infty} s^{-\gamma/2} ds$$
$$= t^{-(\gamma/2-1)}$$

Thus,

$$\sum_{j:d(i,j)>t} G_{ij} \leq M_3 \left(\frac{1}{\delta M_1}\right)^{\gamma/2} t^{-(\gamma/2-1)}$$

This implies that Assumption 6 is satisfied. Next we check Assumption 10.

If $d(k,j), d(k,i) \leq s$ then $\rho(k,j)^2, \rho(k,i)^2 \leq \left(\frac{M_2}{M_3 n}\right)^{2\gamma} s$. Since ρ follows the triangle inequality, then $\rho(i,j)^2 \leq 2\left(\frac{M_2}{M_3 n}\right)^{2\gamma} s$.

Substituting yields $d(i,j) \leq 2\left(\frac{\delta}{\alpha_i}\frac{m_i}{m_j}\right)^{2\gamma}\left(\frac{M_2}{M_3 n}\right)^{2\gamma} s \leq 2\left(\frac{n\delta}{(\alpha+\delta)M_3} M_2\right)^{2\gamma}\left(\frac{M_2}{M_3 n}\right)^{2\gamma} s = 2\left(\frac{\delta}{(\alpha+\delta)M_3} M_2\right)^{2\gamma}\left(\frac{M_2}{M_3}\right)^{2\gamma} s$ which is a constant times s, satisfying Assumption 10.

Next we check Assumption 7. We have that $d(i,j) \leq \left(M_2 n \frac{\delta}{\alpha+\delta}\right)^{2\gamma} \rho(i,j)^2$. We have already assumed that it is possible for every subset Q of the population to construct an s-cover in squared spatial distance $\rho(\cdot,\cdot)^2$ consisting of at most $M_6 \frac{n^{1-2\gamma}}{s}$ units. So an $n^{-2\gamma}\left(M_2 \frac{\delta}{\alpha+\delta}\right)^{-2\gamma} s$-cover in Euclidean distance would contain at most $M_6 \left(M_2 \frac{\delta}{\alpha+\delta}\right)^{2\gamma} \frac{n}{s}$

units. If $\rho(i,j)^2 \leq n^{-\gamma}\left(M_2\frac{\delta}{\alpha+\delta}\right)^{-2\gamma}s$, then $d(i,j) \leq s$ and this must be an s-cover in $d(\cdot,\cdot)$. So we have found an s-cover in $d(\cdot,\cdot)$ containing at most $M_6\left(M_2\frac{\delta}{\alpha+\delta}\right)^{2\gamma}\frac{n}{s}$ units.

Chapter 2, in part is currently being prepared for submission for publication of the material. The dissertation author was one of two primary authors of this material. The other was Professor Paul Niehaus.

Chapter 3

Evaluation of a Teacher Training in Uganda, Specifications for Endline

with Moustafa El-Kashlan, Vesall Nourani, and Sara Restrepo-Tamayo

3.1 Introduction

While 49% of Ugandan children enter secondary school, fewer than 10% enter the final year (Rooke, 2014). In 2007 Uganda eliminated most secondary school fees as part of its Universal Secondary Education (USE) subsidy program. Yet completion rates stagnated three years after USE implementation (World Bank, 2024). Low quality teaching driven by inadequate teacher training may be partly to blame. Although 90% of Ugandan secondary teachers have the required formal qualifications (Vasiliev and Demas, 2018), teachers in schools receiving USE subsidies scored 0.5 standard deviations lower on a numeracy test and 0.28 standard deviations on aa literacy test than teachers in non-implementing schools (Najjumba and Marshall, 2013). Recent reforms to the Ugandan national secondary school curriculum and teacher training standards reflect policymaker awareness of the teaching quality challenge (Businge, 2019).

Kimanya Ngeyo Foundation for Science and Education (Kimanya) was founded in 2007 and has offered a teacher training in primary schools since 2014. The program trains teachers (i) to describe scientific concepts with clarity, (ii) to teach in a participatory

and student-centered manner, and (iii) to reflect on the purpose of education and the importance of what they are doing for their students. These objectives target both general skills and intrinsic motivation. Kimanya training occurs in three rounds of about 11 days each over the course of a school year. In addition, Kimanya tutors visit participating teachers monthly. A randomized evaluation of Kimanya training in primary schools found school-level intention-to-train raised student scores on the high-stakes primary leaving exam by 0.5 standard deviations which raised the pass rate from from 51% to 75% after two years (Nourani et al., 2023).

The present experiment is a separate randomized controlled trial that differs from the first in two key respects.[1] First, the present trial randomizes at both the teacher and school level while (Nourani et al., 2023) randomized only at the school level. The teacher-level design allows us to study teacher-to-teacher spillover effects and also to check whether the program effect remains when teachers cannot easily self-select in training. Second, the present trial takes place in secondary schools and not primary schools. This new setting will allow us to check whether the treatment effect can be replicated in a related but distinct context with older students and a distinct professional class of teachers.

The experiment is ongoing and Figure 3.6 presents the full study timeline. This document is a plan that describes six main research questions that the trial will answer by the end. The plan specifies empirical methods to answer each question. The three datasets needed to answer our questions will be collected at endline in fall 2024 after all treatment teachers have been given the opportunity to train. However, we have already collected midline versions of each of these samples in 2022 and 2023. This document runs our specifications on the midline data in order to demonstrate that our empirical methods are fully specified and sufficiently precise.

We detect no midline effects on student exam scores, possibly because student

[1]AEA registriation is here: https://www.socialscienceregistry.org/trials/12347

exposure to trained teachers was still low or because about 15% of teachers in our study left their jobs prior to 2022 training. However, we do detect changes in teacher attitudes at midline—particularly among teachers who were randomly assigned to train in groups of friends. We will re-run all of these analyses on endline data once it is collected at the end of 2024.

Whatever its findings, this trial will contribute to the literature on improving teaching quality in developing countries. Supply-side interventions often need to come alongside improved teaching to be effective (Glewwe et al., 2009; Kremer et al., 2013; de Ree et al., 2017). Meta-analyses on the impact of teacher training and professional development have yielded promising but highly heterogeneous results (McEwan, 2015; Popova et al., 2021). Recent randomized evaluations find that teachers in developing countries can be quick to effectively adopt innovative new teaching practices (Nourani et al., 2023; Alan and Mumcu, 2024). The present experiment will contribute to understanding of when and how teacher training can improve learning outcomes in developing countries.

The rest of this plan is organized as follows. Section 3.2 specifies six main research questions. Section 3.3 describes our study population, the randomized design, and teacher compliance. Section 3.4 describes our data and discusses balance and attrition. Section 3.5 specifies the empirical methods that we will use to answer each of the six research questions. Section 3.6 runs these methods on midine data to demonstrate that they are fully specified and calculates power for when our methods are re-run on endline data. The remaining sections list the members of the research team, describe our intended deliverables, present the full study calendar, and list our IRB approvals.

3.2 Research Questions

Six primary research questions are enumerated below. Research questions Q1 a, Q1 b, and Q1 c ask whether Kimanya's teacher training program affected student exam scores. Research questions Q2 a, Q2 b, and Q2 c ask whether the training affected teacher attitudes towards gender and the purpose of education. Questions Q1 a, Q1 c, and Q2 a study main effects. Questions Q1 b, Q2 b, and Q2 c study teacher-to-teacher knowledge spillover effects. In Section 3.5.6 we will link each research question to a single primary hypothesis test.

Q1 Student Learning

(a) Can we replicate the impact on student learning outcomes found by (Nourani et al., 2023) in secondary schools? Does the program effect remain when teachers do not self-select into training?

(b) Does training with a friend improve program effectiveness on student exam scores?

(c) What is the program effect on girls' exam scores in particular?

Q2 Teacher attitudes towards the purpose of education for girls vs boys.

(a) Do trained teachers change their perceptions of the differences in ability or the purpose of secondary education between their male and female students?

(b) Does training with a friend magnify the program effect on teacher gender attitudes?

(c) Are untrained teachers' gender attitudes affected by training their colleagues? Does training enough teachers affect school culture?

Several more research questions will be added in the team's final pre-analysis

plan. To meet the deadline of Faridani's dissertation defense, this plan considers only the six questions above that he provided leadership on.

3.3 Research Design

3.3.1 Study Population

Our study consists of 39 secondary schools located in the Jinja and Buikwe districts of southeastern Uganda. We chose the schools in the following fashion. In April-June 2021 we obtained teacher rosters from the 52 secondary schools in both districts and conducted a baseline telephone survey of the teachers on the roster. The survey is described in Section 3.4.1.

We used these survey responses to determine eligibility criteria for teachers. We deemed a teacher to be eligible for training if they (i) taught at least one of the first three years of secondary school and (ii) reported spending at least four hours per week at the school on whose roster they appeared.[2] This left us with 879 eligible teachers. Then we removed the 15 schools with fewest eligible teachers to arrive at 39 schools containing 807 eligible teachers. From these we selected 640 teachers to be part of our randomization.

The set of teachers included in the randomization was chosen to meet two criteria. First, Kimanya wanted to invite approximately 100 teachers per year to train over three years. Second, our research team wanted each school to contribute approximately the same percentage of its eligible teachers to the training, which turned out to be about 60%. We prioritized including teachers who said that they worked more hours on our survey.

Table 3.3 compares teachers not in the randomization to those whom we chose for

[2]Since the survey was taken during the COVID-19 pandemic, four hours worked per week was above the mean in our survey. A national lockdown was implemented on June 7th, midway through our baseline survey.

the randomization. Compared to the other 524 teachers, our chosen 640 teachers were on average 1.4 years older, 8 percentage points more likely to be female, 5 percentage points more likely to be married, 8 percentage points more likely to teach only at the school where they were surveyed, and worked one more hour per week in their surveyed school.

3.3.2 Assignment to Treatment

We randomized at the school and teacher level. The aim of our design was to not only identify the program effect on individual teachers, but also to study teacher-to-teacher spillovers and effects on school culture. We therefore randomize not only who receives training, but also whether teachers are trained in groups of friends or not. To achieve this, the experimental design consisted of four steps: constructing teacher clusters, constructing school triples, school-level randomization, and within-school randomization. Figure 3.1 visualizes the experimental design and the following paragraphs explain the four steps.

The first step in our design was to partition the teachers within each school into four equally-sized treatment clusters using two different methods. For each school we first sorted teachers into what we called *clique clusters*. Clique clusters were chosen in order to maximize the number of teacher-to-teacher social ties within each cluster. Then for each school we sorted those same teachers into four different clusters that we called *anticlique clusters*. Anticlique clusters were chosen in order to minimize the number of social ties within each cluster. The result is that each teacher belonged to one Clique cluster and one Anticlique cluster.

Figure 3.2 visualizes an example of the teacher-level randomization in a small school that sends 2 teachers per year—the minimum and modal number of sent teachers in our sample. The picture visualizes how the teacher social network maps into the Clique and Anticlique clustering schemes. The baseline data used to construct teacher

social networks is described in Section 3.4.1.

In the second step, we sorted the schools into thirteen matched triples of three schools each. Triples were chosen in order to make the number of eligible teachers per school roughly the same within each triple. This was to guarantee that for every realization of the randomization, roughly the same number of teachers would be invited to training.

The third step was to randomize at the school level. Within each triple we randomly selected one school as a Pure Control, one as a Clique school, and one as an Anticlique school. In the 13 Pure Control schools, no one would be invited to train. In the 13 Clique schools, teachers would be invited to train in groups of friends. In the 13 Anticlique schools, teachers would be invited to train in groups with few social ties. Thus our school-level design has three experimental arms: two different treatments and one pure control. We refer to both Clique and Anticlique schools as "treatment schools."

The fourth and final step is to randomize within schools. Each treatment school's teachers were already partitioned into four treatment clusters depending on the school's assignment. In Clique schools we use the four Clique clusters and in Anticlique schools we use the four Anticlique clusters. From each treatment school, one cluster is randomly invited to train in 2022, one cluster is invited to train 2023, one cluster is invited to train in 2024, and the last cluster is kept as a control. See Figure 3.1 for a flowchart.

It is crucial to realize that Clique vs Anticlique assignment has no effect on the conditional probability that any individual teacher is invited to training. Conditioning on Clique vs Anticlique assignment affects only the *joint* distribution of teacher invitation probabilities but not the *marginal* distributions. So, if a teacher-level SUTVA holds where each teachers' outcomes depend only on their own training invitation, then any difference between Clique and Anticlique schools could be interpreted as a violation of SUTVA. This special design allows us to detect teacher-to-teacher spillover effects on

student exam scores even when we cannot link individual students to specific teachers.[3]

Two schools changed ownership between our baseline survey and the second training in April 2023. Consequently both schools experienced very high faculty turnover. Before training began, we re-randomized the 2023 teacher invitations for both schools using the post-ownership-change teacher rosters.

3.3.3 Compliance at Midline

While teacher *invitations* were randomized, actual *attendance* at Kimanya's training is voluntary and therefore cannot be randomized. Moreover, our context includes relatively high teacher employment turnover. To account for imperfect attendance at the trainings, all of the estimands studied later on in this plan are intent-to-treat. In this section we summarize teacher attendances as reported by Kimanya.

While most invited teachers attended significant portions of the training during the first two years of the campaign, some did not. By far the most common reason for non-attendance was that the invited teacher had stopped working at the school that we had initially surveyed them at. Table 3.1 displays all of the compliance information for 2022 and 2023. By the end of 2023, 176 teachers had been invited to training and 131 (75%) of these had attended at least three days of training. Moreover, 79 (45%) of the invited teachers had attended for at least twenty days. Figure 3.3 shows the full CDF of days of training attended for teachers who had been invited by the end of 2023. All 26 treatment schools had sent teachers to train for at least three days by the end of 2023. Among invited teachers, those who attended for at least 3 days were slightly more likely to be female, but did not differ in age from teachers who did not attend training for at least 3 days. Table 3.2 quantifies these comparisons.

The most common reason for noncompliance was that the teacher had stopped

[3]We have attempted to link students to teachers via classrooms, but the links so far are conservative and the resulting networks are quite dense. We therefore do not discuss this kind of analysis in this plan. Our team is working on improving this match and we hope to be able to do this kind of analysis soon.

working at the school where we had initially surveyed them at. Out of 33 invited teachers whom headteachers told us would not attend in 2022 or 2023, for 72% the reason given was that the invited teacher had left the school. This kind of noncompliance will decrease our intent-to-treat effects in a way that is likely unrelated to the impacts of Kimanya training.

In addition, several teachers came to training even though they were not invited. By the end of 2023, 17 uninvited teachers had attended for at least three days and 5 of these had attended at least twenty days.

Teacher attendance patterns suggests that teachers talk to one another about the training and raises the possibility of teacher-to-teacher spillover effects. All 17 uninvited teachers who attended anyway for at least three days worked at treated schools. Nine were from Clique schools and eight from Anticlique.

To avoid wasting capacity, Kimanya wished to minimize the number of empty seats at its trainings due to noncompliance. When a teacher is unavailable or declines an invitation, we invited a replacement teacher from a prepared list. Prior to randomization, for each treatment cluster we prepared an ordered list of "replacement" teachers in the school that were not otherwise part of the randomization. As teachers from a treated cluster declined invitations, we invited replacement teachers in order down the list associated with the cluster of the declining teachers. A total of 65 teachers had been invited from replacement lists by February 2024. Replacement teachers are not included in Table 3.1 or any of the analyses in this plan.

3.4 Data, Attrition, and Balance

This plan discusses five data sources: standardized exams administered by the Uganda National Education Board (UNEB), a teacher survey administered by our team at baseline, an exam given by our team to students, and a midline survey given by our

team to teachers. A similar sample for each dataset will be collected again in Fall 2024. Table 3.4 enumerates all of the datasets discussed in this plan, including future samples. Sections 3.4.1-3.4.4 later on discuss each dataset in detail.

We face significant data attrition in our context. Teachers sometimes change schools or exit the study sample, and one school did not allow its students to sit for our exam. This means that all of our midline and endline datasets are samples, not populations. If treatment causally affects who appears in our data, then this can confound our estimates of treatment effects. We will need to assume that this is not the case.

Assumption 18 states that treatment does not affect which students, teachers, and schools attrit out of our sample. Importantly, Assumption 18 would be violated both by imbalance in the *number* and the *kind* of observations across treatment arms. Without Assumption 18, the estimands studied later on do not have causal interpretations.

Assumption 18. *The sample of units observed in each of our datasets would have been the same under any counterfactual treatment assignment.*

Assumption 18 is not verifiable but we can try to falsify it with balance checks. A balance check tests for differential rates of attrition by treatment arm within each sample. The primary statistic of interest for each balance check is the p-value of a Fisherian exact test, sometimes called "Randomization Inference" (Young, 2018). Here Assumption 18 is the sharp null hypothesis that Fisher's test is testing. The test statistic is always the overall F-score of a regression of treatment status indicators on demographic characteristics. The critical values for the Fisherian exact test are calculated by re-randomizing, computing our test statistic under the counterfactual randomizations, and calculating the percentiles of the distribution of permuted statistics. This test would reject its null hypothesis if, for example, men were over-represented in followup surveys of treated teachers but not control teachers.

The following sections describe how each baseline or midline dataset was col-

lected, provide summary statistics, and check for balance. Assumption 18 is never rejected in any of our samples.

3.4.1 Baseline Teacher Survey

During April-June 2021 our team surveyed 1164 teachers from 54 secondary schools in Jinja and Buikwe. Our survey team was contracted with Innovations for Poverty Action. 834 teachers were surveyed in person and the rest were surveyed over the telephone due to COVID-19 lockdowns. Survey questions included teacher demographics, grades taught, and how many hours they worked at the school that they were surveyed at, and whether they taught at any other schools.

Table 3.3 presents summary statistics for the demographic characteristics of teachers surveyed at baseline. We used this data in November 2021 to select our study population of 640 teachers as described in Section 3.3.1. The summary statistics in Table 3.3 are partitioned into teachers whom we chose to be part of our randomization and teachers whom we did not choose.

We check for balance at baseline by comparing school characteristics across treatment arms. Table 3.5 compares characteristics of treatment schools to pure control schools. Table 3.6 uses the same data to compare Anticlique schools to Clique schools. The Fisherian exact test (randomization inference) fails to reject the null hypothesis of balance at baseline for both comparisons.[4]

The baseline survey was used to get the social networks used to construct the teacher clusters discussed in Section 3.3.2 and illustrated in Figure 3.2. The survey elicited teacher social networks in the following fashion. For each surveyed teacher (called the "ego") we randomly chose fifteen other teachers (called "alters") at the same school. For each alter, we asked the ego whether they had spent 30 minutes talking to

[4]The test statistic for the exact Test is the overall F-score of a regression of school baseline characteristics on school treatment status.

the alter during the last academic term. If the ego answered in the affirmative, then we asked the following three followup questions:

- Have you and [ALTER] spoken about how to better **manage** your classroom?
- Have you or [ALTER] **visited** any of each others classroom to help improve teaching practices?
- Have you **planned** classroom activities together with [ALTER]?

We considered two teachers to be "linked" if they both answered in the affirmative about the other to at least one of the three followup questions. Teacher clusters in the Clique schools described in Section 3.3.2 were chosen to maximize links and clusters in Anticlique schools were designed to minimize links.

3.4.2 Student Assessment proctored by IPA

In October and November of 2022 we gave an exam to 11,495 students. One school (assigned to Anticliques) did not allow us to give the exam, so our exam data comes from 38 schools. The exam was administered by a team contracted with Innovations for Poverty Action. The exam-takers were every student, from S1 to S4,[5] present at the school on the day that the IPA team arrived to give the assessment. The students weren't forced to participate in the assessment but the teacher was present while the students were taking the assessment, increasing the participation of the students on the assessment. Table 3.7 provides summary statistics for the demographics of students who took our assessment.

Our exam consisted of 24 multi-part questions and students had one hour to complete it. A subset of questions were designed bespoke for this project by El-Kashlan and Nourani, while the remaining were taken from the 2011 TIMSS standardized

[5]The first year of secondary school is called S1, the second S2, and so on.

examination, which at the time of study was the most recently available version of the exam available to researchers. The objective of the exam was to measure students' ability to use scientific language and to apply critical thinking skills to scientific problems. The exam questions were similar in spirit to those that the teachers are asked during Kimanya's training sessions. Figure 3.4 provides an example of a question from the "Use of Language" section where students are asked to decide when to use the word "science" and when to use the word "technology".[6] Figure 3.5 shows an example of a question from the "Critical Thinking" section where students are asked to construct a simple experiment to test a hypothesis.

We computed exam scores by taking 35 question-parts that had single correct answers and adding up the number of correct answers per exam. Then we subtracted the mean and divided by the standard deviation over all exams.

Since one Anticlique school did not take our assessment, we are concerned that student exams might not be balanced between treatment and control. Table 3.7 compares the demographic characteristics of students who took our assessment across treatment and control schools. We check for balance in the following characteristics: student age, student gender, indicators of whether the student was in year 2, 3, or 4 of secondary school, and the mean UCE score for the student's school prior to baseline.

To check for balance overall, we run a Fisherian exact test using the overall F-score of a regression of the treatment indicator on the six characteristics in Table 3.7 as our test statistic. The test fails to reject Assumption 18.

We gave a similar exam in Fall 2023 but has not been digitized yet and so we cannot analyze it here. We will give a third exam in Fall 2024.

[6]Since the S4 students were preparing for their UCE exam, we solely administered Section 1 - Use of Language to them.

3.4.3 UNEB Exam Scores

Students in the fourth year of secondary school (called S4) take the Uganda Certificate of Education (UCE) exam every year. The exam is administered by the Uganda National Board of Education (UNEB). The exam is mandatory and high-stakes and is comparable to the "O-level" exam in the United Kingdom.

The UCE exam contains six mandatory subjects: Mathematics (150 minutes), Biology (120 minutes), English language (120 minutes), Chemistry (120 minutes), Physics, (120 minutes), and either Geography (150 minutes), or History (120 minutes) or Religious education (90-120 minutes).[7] Scores are aggregated across all subjects by UNEB. We process aggregate scores by subtracting the mean and dividing by the standard deviation.

We attempt to match student's UCE exam scores to their score on the Primary Leaving Exam (PLE). The PLE is a high-stakes exam taken in the last year of primary school that is used for admissions into secondary school. We use standardized PLE scores as a student-level covariate in our analyses because they were determined prior to treatment and are strongly correlated with UCE exam scores. Similar to the UCE, our PLE scores have been aggregated across all subjects by UNEB.

UCE and PLE scores were provided and matched by UNEB. We obtained data by requesting that UNEB provide scores for exams taken by students attending schools in our study area. Unfortunately, UNEB's data is linked not to students' schools but only to the testing center that serviced that school. To solve this problem, we asked schools to provide the index numbers and testing centers of exams taken by their students. We then use students' names to improve the match.

We have matched data from 2018-2020 and 2022. The exams prior to baseline are processed differently than exams after baseline, so we will discuss them separately.[8]

[7] http://nada.uis.unesco.org/nada/fr/index.php/catalogue/47

[8] While we do have 2023 data from UNEB in hand, we need to obtain more information from schools to

UCE Prior to Randomization

Our data contain 12731 UCE exams taken at 49 testing centers during 2018-2020. These testing centers were selected to correspond to the schools in our randomization. Since these exams were taken before randomization we treat them as pre-determined covariates. We were able to match 11483 out of 12731 pre-baseline exams to 37 out of our 39 schools. One missing school was assigned to Cliques and the other to Anticliques.

Since not every exam was successfully matched to a school, we may be concerned that Assumption 18 was violated somehow. We do not find evidence that the number of observations differed by school treatment status. 3623 scores were from Clique schools, 4080 from Anticlique schools, and 3780 from Pure Control Schools. Fisher's exact test does not reject the null hypothesis that the number of exams per school was unaffected by school treatment status.

We use these pre-baseline exam scores as covariates in our analyses. We do this by taking the mean of standardized UCE exam scores within each school-year and then meaning the school-years within each school. For the two schools without any baseline scores, we impute their mean standardized scores as zero. Imputation does not affect the validity or interpretation of our analyses because these means are used as predetermined covariates in the doubly robust estimators described in Section 3.5.5. Balance remains even after imputation. School mean standardized scores are only .04 standard deviations higher in treated schools than control schools. Fisher's exact test finds that this difference is statistically insignificant.

UCE Scores in 2022

Our data contain the universe of exams in Buikwe and Jinja districts taken in 2022, which is 14,595 exams.[9] Since these exams were taken post-baseline we treat

complete the match. We anticipate receiving 2024 data from UNEB in March 2025.

[9]We have data from a significantly larger area in 2022 than in previous years. This makes the sample size of matched data is somewhat larger (4456 in 2022 vs 3782 in 2020).

2022 scores as an outcome variable. We were able to match 4456 exams to students in 38 out of 39 of the schools in our randomization. The missing school did not provide index numbers for their students in 2022.

The number of matched exams does not seem to differ by treatment arm. Our matched data contain 1462 exams from Clique schools, 1436 from Anticlique, and 1558 from Pure Control. These do not differ significantly from the 1485 exams per treatment arm that we would expect under perfect balance.

Table 3.8 shows balance for our sample of 2022 UCE test-takers. Treatment and control schools are very similar in this sample across all pre-baseline characteristics. The Fisherian exact test finds no evidence against Assumption 18.

We have UCE exam scores from 2023 but the data has not been fully matched to schools in our randomization and is therefore not analyzed in this plan. We anticipate to receive 2024 exam scores from UNEB in March 2025.

3.4.4 Teacher Survey

In October and November of 2023 our team surveyed 1153 teachers at all 39 schools. We attempted to reach all 640 teachers in the randomization but only reached 455. The other 698 survey responses were from teachers not in the randomization and will not be discussed further in this document.

The key outcome variables discussed in this document are the responses to one particular survey module that asks teachers the extent to which they agree with each of eight statements about how gender shapes the purpose of education. The purpose of these questions is to determine whether training changes teachers' attitudes towards boys' and girls' education. The statements are listed below. Table 3.9 provides summary statistics.

1. Boys have more ability than girls to learn science subjects.

2. Boys need more education because they are better suited than girls to find high-paying jobs.

3. Parents should maintain stricter control over their daughters than their sons.

4. Girls should attain higher education so that they find better husbands.

5. Boys should attain higher education so that they find better wives.

6. Boys should be just as capable as girls at preparing a meal for visitors when guests visit.

7. Girls have more ability than boys to learn science subjects.

8. Girls need more education because they are better suited than boys to find high-paying jobs.

Since 185 teachers from our randomization are missing, we must test whether this attrition violated Assumption 18. Our first piece of evidence in favor of Assumption 18 is that 65% of survey respondents come from treated schools which is very close to the 66% of the study population of 640 teachers were in treatment schools. Second, 53% of survey respondents had been invited to training by 2023 which is very close to the 50% of the study population that was assigned to receive an invitation. This suggests that treatment had no influence on the number of survey responses.

We next wish to check that treatment did not change who responded. Table 3.10 compares characteristics of survey respondents across pure control and treated schools while Table 3.11 compares characteristics of respondents who had been invited by 2023 to uninvited respondents. We do not find statistically significant differences for any one characteristic. To check for balance overall, we run the Fisherian exact test using the overall F-score of a regression of the treatment indicator on the teacher characteristics as

our test statistic. Fisher's test finds no evidence against the null hypothesis of Assumption 18.

We plan to collect another round of teacher survey data in October and November of 2024.

3.5 Emprical Methods

3.5.1 Primary Outcome Variables

We study effects on three primary outcome variables:

- Research questions Q1 a, Q1 b, and Q1 c consider student scores on two exams:
  - The fraction of correct responses on our student exam described in Section 3.4.2.
  - All-subject aggregate score on the Uganda Certificate of Education standardized exam described in Section 3.4.3.
- Research questions Q2 a, Q2 b, and Q2 c consider indicators of whether teachers agreed with eight gender-related statements on our teacher survey described in Section 3.4.4.

3.5.2 Notation

Consider a sample of m teachers and n students who appear in our data. Let T_{jts} be the treatment assignment of teacher j at school s in year t and let $\mathbf{T}$ be the full $3m \times 1$ vector of all teacher treatment assignments for all three years of the study. Let $Y_{its}(\mathbf{T})$ be the potential exam score for student i in year t of the study attending school s under teacher treatment assignment vector $\mathbf{T}$. Let $D_s(\mathbf{T})$ be the assignment of school s (Clique, Anticlique, or Pure Control). For brevity we will sometimes suppress the argument $\mathbf{T}$.

In our framework, potential outcomes are fixed and treatment assignment is random. This means that expectations are taken over all treatment assignments and not

over any sampling procedure. This requires Assumption 18 to hold, i.e. that treatment assignment does not change who is observed.

The following sections describe our estimands and hypotheses. These are subscripted according to the following convention. The first part of the subscript corresponds to the research question that the estimand or hypothesis is meant to address, and the second part of the subscript is a unique identifying number. So $\Theta_{Q1b.3.2}$ is a treatment effect that addresses question Q1 b. Clicking on the first part of the subscript will take the reader to the specification of the research question. Clicking on the second part of the subscript will take the reader to the definition of the estimand.

3.5.3 ITT Effects on Student Exam Scores

The following treatment effects attempt to address the main research questions Q1 a, Q1 b, and Q1 c by studying treatment effects on student exam scores Y_{its}. We will study effects on both our own assessment that we give to the students described in Section 3.4.2 and the UCE exam described in Section 3.4.3.

All of these are intent-to-treat effects, meaning that they are the causal effects of receiving an invitation to train, not the effect of attending training. Recall that in our framework treatment assignment is random, but under Assumption 18 the sample is fixed and non-random. So Assumption 18 guarantees that all of the treatment effects in this section are identified even when some units from the randomization do not appear in the sample. Take note however that these are effects only on the (fixed) sample of units who appear in our data.

The estimand $\Theta_{Q1a.3.1}$ captures the school-level causal effect of receiving training on student test scores. It most closely resembles the effect estimated by Nourani et al. (2023). It compares assessment scores across treatment vs control schools within a single

year t.[10]

$$\Theta_{Q1a.3.1} \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_{its} \,|\, D_s \neq \text{CTL}\right] - \mathbb{E}\left[Y_{its} \,|\, D_s = \text{CTL}\right] \tag{3.1}$$

The next estimand studies teacher-to-teacher knowledge spillovers. Estimand $\Theta_{Q1b.3.2}$ leverages the special design of the experiment to compare schools where teachers who trained with a friend (Clique schools) to schools where those who did not (Anticlique schools) within a given year t. If we find that $\Theta_{Q1b.3.2} \neq 0$ then we will have found evidence of teacher-to-teacher spillover effects on students without needing to link students to teachers—a difficult data task.

$$\Theta_{Q1b.3.2} \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_{its} \,|\, D_s = \text{CLIQUE}\right] - \mathbb{E}\left[Y_{its} \,|\, D_s = \text{ANTI}\right] \tag{3.2}$$

The third estimand $\Theta_{Q1c.3.3}$ captures the school-level causal effect of receiving training on student test scores for female students only. It compares girls' scores across treatment vs control schools within a single year t.

$$\Theta_{Q1c.3.3} \equiv \frac{\sum_{i=1}^{n} \mathbf{1}\left\{\text{FEMALE}_i\right\}\left(\mathbb{E}\left[Y_{its} \,|\, D_s \neq \text{CTL}\right] - \mathbb{E}\left[Y_{its} \,|\, D_s = \text{CTL}\right]\right)}{\sum_{i=1}^{n} \mathbf{1}\left\{\text{FEMALE}_i\right\}} \tag{3.3}$$

3.5.4 ITT Effects on Teacher Survey Responses

The next three estimands study the effect of invitation to training on teacher attitudes towards the purpose of secondary education for boys vs girls. They each correspond to one of the research questions Q2 a, Q2 b, and Q2 c. Let Z_{itsu} be a dummy variable indicating that teacher i at school s responded with agreement to the gender statement u on the teacher survey in year t. The survey questions are described in Section 3.4.4.

The estimand $\Theta_{Q2a.3.4.u}$ compares the rate of agreement among invited teachers

[10]The expectation is over randomized teacher invitations within each school, not over a sampling process.

to teachers in pure control schools. By excluding uninvited teachers in treatment schools we avoid diluting the treatment effect with spillovers.

$$\Theta_{Q2a.3.4.u} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Z_{itsu} \,|\, T_{jts} = 1\right] - \mathbb{E}\left[Z_{itsu} \,|\, D_s = \text{CTL}\right] \tag{3.4}$$

The estimand $\Theta_{Q2b.3.5.u}$ compares average agreement rates for question u for invited teachers across Clique and Anticlique schools. This captures the effect of training with friends vs training with co-workers who are not friends.

$$\Theta_{Q2b.3.5.u} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Z_{itsu} \,|\, D_s = \text{CLIQUE} \cap T_{jts} = 1\right] - \mathbb{E}\left[Z_{itsu} \,|\, D_s = \text{ANTI} \cap T_{jts} = 1\right] \tag{3.5}$$

The estimand $\Theta_{Q2c.3.6.u}$ compares average agreement rates for question u by uninvited teachers across treatment and pure control schools. This comparison captures the effect of being exposed to trained colleagues on teachers who were not themselves invited to train.

$$\Theta_{Q2c.3.6.u} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Z_{itsu} \,|\, D_s \neq \text{CTL} \cap T_{jts} = 0\right] - \mathbb{E}\left[Z_{itsu} \,|\, D_s = \text{CTL}\right] \tag{3.6}$$

3.5.5 Estimation

We estimate Θ in Sections 3.5.3-3.5.4 with the regression-adjusted estimator from Wooldridge (2010) on pp. 930-931. The relevant STATA command is: "teffects ipwra." We describe here the estimation procedure for $\Theta_{Q1b.3.2}$, but the procedure is nearly identical for all of the estimands. There are two steps. In the first step we run the OLS regression in Equation 3.7. No weights are necessary because the true propensity scores are all constant.

$$Y_{its} = 1\left\{D_s = \text{CLIQUE}\right\} X_{is}\beta_1 + 1\left\{D_s \neq \text{CLIQUE}\right\} X_{is}\beta_0 + \varepsilon_{its} \tag{3.7}$$

To increase precision, we include the $k \times 1$ vector of baseline covariates X_{is}. These always include a constant, pre-baseline UCE exam scores meaned within each school, student (or teacher) age, student (or teacher) gender, and (when estimating effects on students) student cohort indicators. When we estimate effects on UCE scores we also include student PLE scores as an additional covariate. The regression in Equation 3.7 predicts outcomes using the covariates separately among students in Clique schools and those in Anticlique schools. The regression yields the two $k \times 1$ vectors coefficients $\hat{\beta}_1, \hat{\beta}_0$.

In the third step, we compute the final estimate by taking the difference in the predicted values of Y_{its} under the two treatment conditions.

$$\hat{\Theta}_{Q1b.3.2} = \frac{1}{n} \sum_{i=1}^{n} X_{is} \left(\hat{\beta}_1 - \hat{\beta}_0 \right)$$

In order for $\hat{\Theta}_{Q1b.3.2}$ to be consistent for $\Theta_{Q1b.3.2}$, we must verify the overlap and unconfoundedness conditions. The experimental design guarantees that every student and teacher has a chance to be in any group that the estimands compare, so overlap is satisfied. Since all treatments were randomly assigned, if we could observe data from every teacher and student at every school, unconfoundedness would hold whenever we condition on covariates pre-determined at baseline. Since we only observe a sample of teachers and students, we need Assumption 18 to hold in order to have unconfoundedness.

We use the regression-adjusted estimator because it has the double robustness property. Since we know the true treatment probabilities of each student, school, and teacher, our propensity score model is true by design. This means that we can adjust for any covariates using any model so long as the covariates are determined prior to randomization and Assumption 18 holds. We choose to specify a linear model for the outcomes in Equation 3.7. Even if the linear model is false, we can still consistently estimate all of our estimands because of double robustness.

Adjusting for covariates is vital in our case because they greatly increase precision by explaining much of the variation in our outcome variables. For example, student PLE scores from primary school explain 44% of the UCE exams scores taken in secondary school. The doubly robust estimator allows us to exploit these powerful covariates without imposing any additional assumptions on how outcomes are determined.

3.5.6 Primary Hypotheses

While we will test many hypotheses, we specify only one "primary" test for each of our six research questions from Section 3.2. These tests are enumerated in Table 3.12. All six primary tests will be tested exclusively using outcome data from 2024—which has not been collected at the time that this report is written. The table makes explicit the sharp null hypothesis being tested, provides an interpretation in plain English, and notes the test statistic to be used. We do not adjust for multiple hypothesis testing across the six distinct research questions because they are of separate interest. Since each research question is associated with only one primary test (sometimes a joint test), there is no need to adjust for multiple testing within any research question.

The first three primary tests in Table 3.12 check for effects of training on student exam scores. The primary null hypothesis for Q1 a is that school treatment status will have no effect on any student's score on the 2024 UCE exam. The primary null hypothesis for Q1 b is that school assignment to Clique vs Anticlique will have no effect on any student's score on the 2024 UCE exam. The primary null hypothesis for Q1 c is that school treatment status had no effect on any female student's score on the 2024 UCE exam.

The last three primary tests in Table 3.12 consider effects on teacher attitudes toward gender and the purpose of education for girls and boys. The primary test for Q2 a tests the sharp null hypothesis that receiving an invitation to training did not affect any teacher's responses to any of the eight survey questions in our gender module from

Section 3.4.4. The primary test for Q2 b tests the sharp null hypothesis that no teacher's responses to the gender module were affected by whether they were invited to train alongside friends or not. The primary test for Q2 c tests the sharp null hypothesis that no uninvited teacher's responses to the gender module were affected by whether any other teachers at their school had been invited to train.

3.5.7 Hypothesis Testing Methods

Next we describe how each of these sharp null hypotheses will be tested. We will not use t-tests for inference because the number of treatment clusters is small. While we report robust standard errors clustered at the school level for all estimates $\hat{\Theta}$ discussed above, these standard errors are unreliable because the number of clusters (38 or 39 schools) is small.

Our solution is to exploit the randomized design to use a Fisherian exact test known to economists as "randomization inference" (Young, 2018). This procedure only tests the "sharp" null hypotheses under which all outcomes used to construct a particular test statistic are known. Table 3.12 explicitly lists each sharp null that is being tested. The validity of Fisher's test requires only Assumption 18, i.e. that treatment never causes (non)attrition from the sample. Fisher's test is exact and requires no other assumptions or asymptotic approximations and will control the type-I error rate regardless of the number or size of the treatment clusters.

Table 3.12 lists the statistics that we will use to test each research question. For the three research questions concerning student outcomes Q1 a, Q1 b, and Q1 c we specify the test statistics to be the point estimates of the treatment effects, $\hat{\Theta}_{Q1a.3.1}$, $\hat{\Theta}_{Q1b.3.2}$ and $\hat{\Theta}_{Q1c.3.3}$ respectively.

The research questions Q2 a, Q2 b, and Q2 c consider several outcome variables each. To produce a single test statistic per hypothesis, we use the following three F-scores. Define the statistic $F_{Q2a.3.8}$ in the following way. First run the OLS regression

in Equation 3.8 below. Exclude all teachers in treatment schools who had not yet been invited to train in year t. Let T_{jts} indicate that teacher j at school s had been invited to train by year t. Let Z_{jstu} indicate that the teacher agreed with the statement in question u. Include that teacher's baseline age and gender, as well as school s mean UCE score prior to randomization as covariates. Define $F_{Q2a.3.8}$ as the F-score from the usual F-test of the null hypothesis that every γ_u coefficient is equal to zero where standard errors were clustered at the school level. This F-score will be large when teacher gender question responses predict whether the teacher had been invited to training or was in a pure control school.

$$F_{Q2a.3.8}: \quad T_{jts} = \beta_0 + \beta_1 \text{AGE}_{js} + \beta_2 \text{MALE}_{js} + \beta_3 \text{UCE}_s + \sum_u \gamma_u Z_{jstu} + \varepsilon_{jts} \tag{3.8}$$

To test the null hypothesis for Q2 b, define $F_{Q2b.3.9}$ as the F-score from the usual F-test of the null hypothesis that every γ_u coefficient is equal to zero in Equation 3.9 below. Teachers in pure control schools are excluded. This F-score will be large when gender question answers predict whether the teacher was invited to train in a Clique or an anticlique.

$$F_{Q2b.3.9}: \quad \mathbf{1}\{D_s = \text{CLIQUE}\} = \beta_0 + \beta_1 \text{AGE}_{js} + \beta_2 \text{MALE}_{js} + \beta_3 \text{UCE}_s + \sum_u \gamma_u Z_{jstu} + \varepsilon_{jts} \tag{3.9}$$

Finally, to test the null hypothesis for question Q2 c, define $F_{Q2c.3.10}$ as the F-score from the usual F-test of the null hypothesis that every γ_u coefficient is equal to zero in Equation 3.10 below. Teachers who were invited to training are excluded. This F-score will be large when gender question answers among teachers who were never invited to train nevertheless predict whether their school was assigned to pure control.

$$F_{Q2c.3.10}: \quad \mathbf{1}\{D_s \neq \text{CTL}\} = \beta_0 + \beta_1 \text{AGE}_{js} + \beta_2 \text{MALE}_{js} + \beta_3 \text{UCE}_s + \sum_u \gamma_u Z_{jstu} + \varepsilon_{jts} \tag{3.10}$$

3.6 Midline Results and Endline Power

We will test the six primary hypotheses in Table 3.12 at endline once the 2024 teacher survey and student exam scores are collected. To demonstrate that the testing and estimation methods are fully specified, we present here results using midline data collected at the end of 2022 and 2023. We detect no midline effect on student exam scores, likely because student exposure to trained teachers at midline was still low. We do detect midline effects on teacher gender attitudes. This effect is significantly larger in Clique schools, implying teacher-to-teacher spillovers. Midline results are not used to draw final conclusions about the intervention, but they do provide a full set of specifications for endline.

3.6.1 Effects on Student Exam Scores

Table 3.13 shows estimates of $\Theta_{Q1a.3.1}$ which compares student exam scores in treated to untreated schools. Table 3.14 shows estimates of $\Theta_{Q1b.3.2}$ which compares student exam scores in Clique to Anticlique schools. Table 3.15 shows $\Theta_{Q1c.3.2}$ which is the treatment effects for girls. The columns of each table show the point estimate, cluster-robust standard error, and the p-value of the Fisherian exact test that uses the point estimate as the test statistic. The top row of each table uses data from our 2022 student assessment described in Section 3.4.2 and the bottom row uses the 2022 UCE exam scores described in Section 3.4.3. While the point estimates are nearly always positive, Fisher's test does not reject any of the three null hypothesis on either of the two exams.

The null midline effect on exam scores is unsurprising for two reasons. First, Nourani et al. (2023) also did not find program effects on exam scores after only one year. Second, when students took our assessment in 2022, the first cohort of teachers were still going through Kimanya training.

We expect larger effects on exam scores at endline. By the end of 2024, three times as many teachers will have been trained as 2022, the treated teachers will have been trained for longer, and therefore more students will have had more years of exposure to trained teachers. We calculate power to detect a larger effect in the following way. We simulate endline samples by resampling the midline data with replacement within each school and adding a hypothetical treatment effect to the resampled scores. Power is calculated as the fraction of bootstrapped samples in which our tests reject their sharp null hypotheses at the 5% level. Resampling in this way overstates within-cluster correlations and is likely to understate true power.

The right panel of Table 3.13 shows the results of the power calculations. Here the null hypothesis is that school assignment to treatment or pure control had no effect on any exam score. We calculate at least 92% power to detect an endline effect .38 standard deviations larger than midline on our assessment and an effect .24 standard deviations larger than midline on the UCE exam.

The right pane of Table 3.14 shows power to reject the sharp null hypothesis that Clique and Anticlique treatment have the same effect on exam scores for all students. We find 90% power to detect a Clique effect of .38 standard deviations on scores for our student assessment and 95% power to detect an effect of .19 standard deviations on the UCE scores.

The right pane of Table 3.15 shows power to reject the sharp null hypothesis that treatment affected girls' exam scores. We find at 93% power to detect a Clique effect of .25 standard deviations on scores for our student assessment and 95% power to detect an effect of .39 standard deviations on the UCE scores. We conclude that fir girls' exams, our assessment is much more likely to detect an effect than for UCE exams.

These detectable effect sizes are significantly smaller than the effects found by Nourani et al. (2023) of 0.5 standard deviations on high-stakes exam scores. Moreover these power calculations are likely to understate true statistical power because resampling

within schools overstates within-cluster dependence for small schools. Finally, these effect sizes, while still large, are detectable with very high probability.

3.6.2 Effects on Teacher Attitudes

We do find significant effects at midline on teacher attitudes towards the purpose of education for boys vs girls. Table 3.16 shows estimates of $\Theta_{Q2a.3.4}$, which compares trained teachers to pure control teachers. While Fisher's test fails to reject the null hypothesis for any individual survey question, the joint null hypothesis of no effect on any question is rejected with $p = .01$. We are interested to see whether the endline survey yields more insight but not much can be said based on the midline data.

Table 3.17 shows much stronger effects. Here we estimate $\Theta_{Q2b.3.5}$, which compares invited teachers in Clique schools to invited teachers in Anticlique schools. The joint null hypothesis of no effect on any of the eight survey questions is rejected with Fisher's $p = .02$. In particular, we find that trained teachers from Clique schools are significantly less likely to agree with each of the following three statements:

- Boys need more education because they are better suited than girls to find high-paying jobs. ($p = .00$)
- Girls should attain higher education so that they find better husbands. ($p = .00$)
- Girls need more education because they are better suited than boys to find high-paying jobs. ($p = .04$)

The fact that midline effects on Clique vs Anticlique invited teachers are larger than effects on invited vs pure control teachers leads us to hypothesize that training with a friend is more effective than not training with a friend. We look forward to confirming this result with the endline survey.

We do not find evidence of spillovers onto untrained teachers at midline. Table 3.18 shows estimates of $\Theta_{Q2c.3.6}$ which compares uninvited teachers in treatment vs

control schools. Neither the joint test nor the individual tests reject their sharp null hypotheses of no spillover effect of having trained colleagues on untrained teachers. Since the joint p-value of .12 is relatively low, we are interested to find out whether spillover effects appear at endline when more teachers are trained.

We wish to calculate power to reject the three joint null hypotheses in Table 3.12 at endline. In contrast to students, we will not collect a new sample of teachers in 2024 but rather re-survey the same set of teachers as before. To calculate power, we ask two questions. First, how much would teacher responses have to change in order for us to fail to detect effects on already seen at midline? Second, how much would the spillover effects on untrained teachers in 2024 have to be for us to detect it at midline? In other words, we run power calculations that condition upon midline survey responses.

Power calculations suggest that the finding in Table 3.16 might be difficult to replicate. In a simulation we randomly chose 15% of teacher-questions and changed their survey responses at random. Then for two survey questions, we changed 35% of "Agrees" among teachers trained in or before 2024 to "Disagrees". In 76% of simulations, Fisher's test using $F_{Q2a.3.8}$ still rejected the null. This suggests that if even a modest amount of noise is added to teacher responses, then the program effect in year 3 would need to be very strong for us to replicate the p-value at the bottom of Table 3.16.

In contrast, conditional power calculations suggest that we are very likely to replicate the rejection of the joint null hypothesis in Table 3.17 even if survey responses are noisy. In a simulation we randomly chose 15% of teacher-questions and changed their survey responses at random. But here we did not add any program effects at all for teachers newly trained in 2024. Fisher's test still rejected the joint null hypothesis in 98% of simulations. This suggests that a contrast between Clique and Anticlique schools is likely to hold up at endline—even if the endline survey is noisy and third year program effects are small.

While Table 3.18 does not show midline effects on uninvited teachers, we wish to

know what magnitude of effect is likely to be detectable at endline. A simulation exercise showed that if every teacher-question answered by a teacher still uninvited in 2024 has a 40% chance to change from agree to disagree, then Fisher's test using $F_{Q2c.3.10}$ only rejects the sharp null hypothesis 66% of the time. This is because only one quarter of teachers in treatment schools will remain uninvited in 2024. Although the true spillover effect is likely highest when most teachers are trained, the sample used to detect that spillover is at its smallest when most teachers are trained. Power to detect an endline effect on untrained teachers will depend on whether the spillover has become large by that time.

3.7 Research Team

The lead principal investigator is Vesall Nourani. Nourani is responsible for coordinating the study and building relations with Kimanya, the Echidna Foundation, and other partners. Moustafa El-Kashlan and Sara Restrepo-Tamayo lead data collection and survey instrument design. Stefan Faridani developed the experimental design and leads the quantitative data analysis. All four will contribute as co-authors to writing academic papers and other deliverables. We are greatful for the field and data assistance given by Faith Babirye and Jiya Nair's research coordination.

3.8 Deliverables

We expect the data and analysis from this experiment to result in two articles published in academic journals. The authors for both articles will be El-Kashlan, Faridani, Nourani, and Tamayo. The first article will study spillover effects of the training across teachers and students. The second article will study how training affects teachers' thoughts and behavior. In particular the second article will focus on the alignment between teacher's stated purpose for each pedagogical technique that they use and

students' perceived purpose of the pedagogical techniques that their teachers use. Both articles will consider how changes in the Ugandan national secondary school curriculum interacts with training to affect teachers practices and what they report the purpose of those practices to be. We expect that both articles will be submitted to academic journals in Fall 2025.

3.9 Calendar

The timeline in Figure 3.6 shows all completed and upcoming tasks. The most important dates are the following. Baseline surveys began in May 2021 and endline surveys are scheduled to be complete in August 2024. Randomization occurred on November 29, 2021 and teacher training consisted of three waves beginning in April 2022, December 2022, and December 2023.

3.10 TIMSS Citation

Several questions on our secondary school assessment described in Section 3.4.2 used modified versions of questions from the TIMSS 2011 exam. While these TIMSS questions are in the public domain, it was requested that we include the following special citation:

```
SOURCE: TIMSS 2011 Assessment. Copyright © 2013 International Association
for the Evaluation of  Educational Achievement (IEA). Publisher: TIMSS &
PIRLS International Study Center, Lynch School  of Education, Boston
College, Chestnut Hill, MA and International Association for the
Evaluation of Educational Achievement (IEA),
IEA Secretariat, Amsterdam, the Netherlands
```

3.11 Figures for Chapter 3

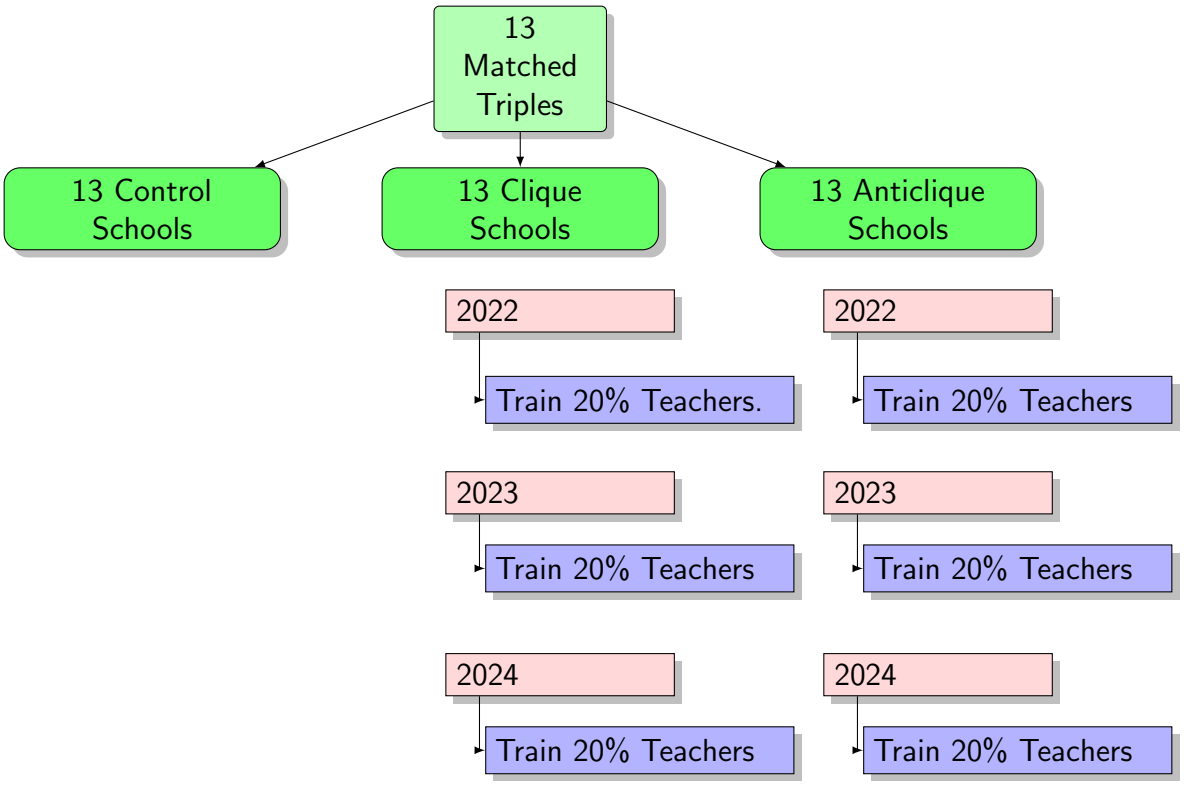

Figure 3.1. School-Level Design Flowchart.

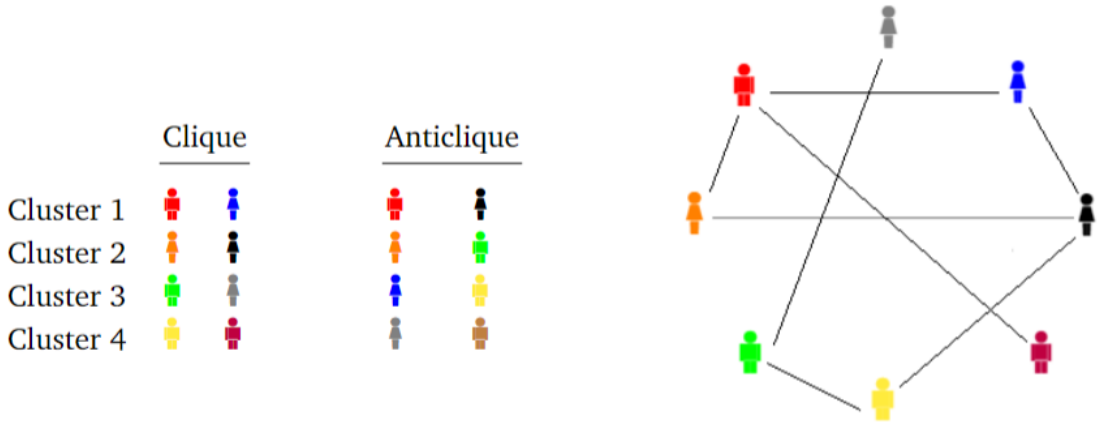

Figure 3.2. Example clustering of teachers. The left panel shows the clustering scheme and the right panel shows the social network.

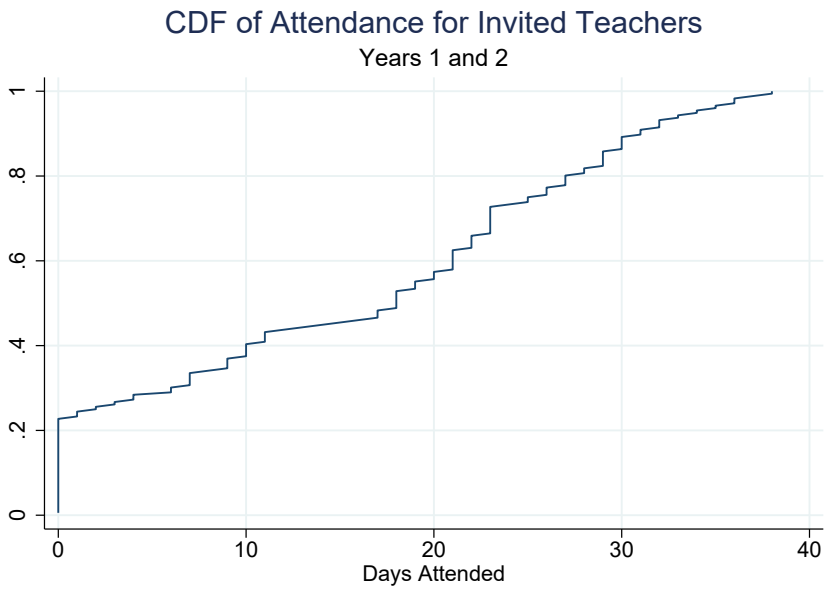

Figure 3.3. CDF of Teacher Attendance by end of 2023

2. Choose one of the following option choices to fill each of the sentences **a** to **d** that follow:

   OPTION CHOICES: ***science, technology, both, neither.***

   a. __science__ helps us discover how diseases operate.

   b. Medicine is a __science(full-credit)/both(half-credit)__ that helps us prevent disease.

   c. Vaccines are a _technology____ that help us prevent diseases.

   d. The __science___ of nutrition can help us maintain good health.

Figure 3.4. Example of questions from the Use of Language section of the 2022 student assessment with correct answers highlighted in yellow.

1. Susie has a potted plant. She sets up an experiment that shows that water travels through a plant into the air.

Which experiment would show this?[☑ the one correct answer]

☐ **A.** Put water in a container under the pot; water will disappear from the container.

☐ **B.** Cover one of the stems of the plant with a plastic bag and water the plant; drops of water will be seen in the bag.

☐ **C.** Place a cut stem from the plant in a plastic bag; water will be seen in the bag.

☐ **D.** Place a cut stem from the plant in a glass of colored water; the plant's leaves will change color.

Figure 3.5. Example of questions from the Critical Thinking section of the 2022 student assessment. The correct answer is B.

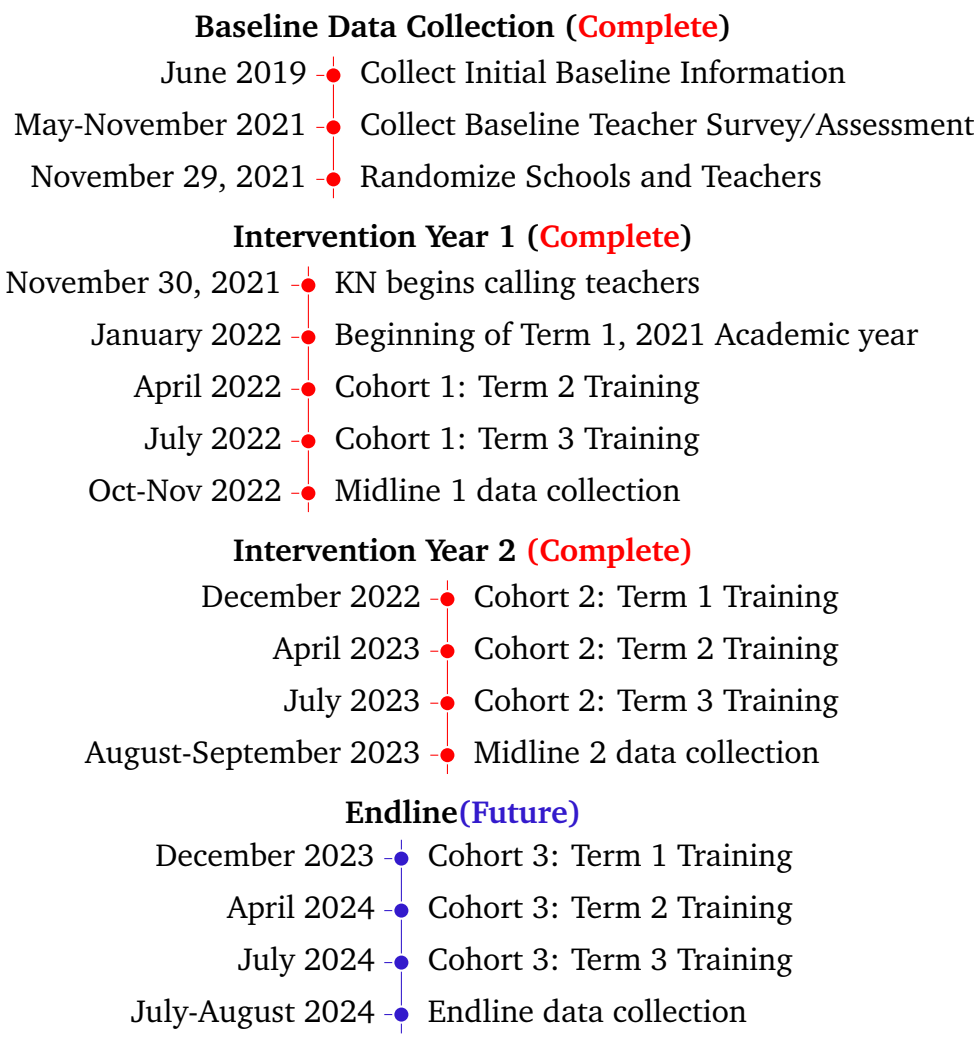

Figure 3.6. Study Timeline

3.12 Tables

Table 3.1. Compliance Tabulations.

	Compliance by end 2022			Compliance by end 2023		
	Not Inv. '22	Invited '22	Total	Not Inv. '23	Invited '23	Total
$<$ 3 days	549	24	573	447	45	492
$\geq$ 3 days	6	61	67	17	131	148
$<$ 20 days	549	42	591	459	97	556
$\geq$ 20 days	6	43	49	5	79	84
Total	555	84	640	464	176	640

Notes: Columns divide teachers by intention to treat in 2022 (left panel) and 2023 (right panel). Rows show number of teachers who actually attended more than 3 days (or 20 days) of Kimanya training.

Table 3.2. Differences between non-attendees (left) and attendees (right).

Demographics of Invitees by Compliance			
	(Attended $<$ 3 days)	(Attended $\geq$ 3 days)	(Diff)
Teacher Age	40.233	38.714	-1.518
	(10.851)	(10.388)	(2.094)
Teacher Male	0.814	0.683	-0.131*
	(0.394)	(0.467)	(0.066)
Observations	45	131	176

Table 3.3. Summary statistics for teachers at baseline.

	(Not)	(In Randomization)	(Difference)
Teacher Age	37.332	38.689	1.357**
	(9.976)	(10.045)	(0.590)
Teacher Male	0.761	0.678	-0.083***
	(0.427)	(0.468)	(0.026)
Teacher Married	0.714	0.769	0.055**
	(0.452)	(0.422)	(0.026)
Teacher holds Undergraduate	0.771	0.789	0.018
	(0.421)	(0.408)	(0.024)
Hours worked per week here	3.403	4.394	0.991***
	(2.906)	(1.496)	(0.132)
Teach at other schools?	0.189	0.106	-0.083**
	(0.393)	(0.309)	(0.039)
Teachers	524	640	1,164
Schools	53	39	54

Notes: Standard deviations are in parentheses in the first columns. Standard errors are in parentheses in the third column. Hours worked per week were low due to COVID-19.

Table 3.4. Summarizes datasets discussed in this plan—present and future.

Dataset	Section	Unit	Year	Collected & Clean?	*N*
UCE Exam	3.4.3	Student	'18-'20	✓	12474
UCE Exam	3.4.3	Student	'22	✓	4456
UCE Exam		Student	'24		
Assessment	3.4.2	Student	'22	✓	11495
Assessment		Student	'23		
Assessment		Student	'24		
Baseline	3.4.1	Teacher	'21	✓	640
Teacher Survey	3.4.4	Teacher	'23	✓	455
Teacher Survey		Teacher	'24		

Notes: The second column displays the section of this document that describes the sample. The final column reports the number of observations that are non-missing and linked to a member of our randomization. Note: the '23 student assessment has been collected but not yet digitized.

Table 3.5. School Balance at Baseline: Treated vs Pure Control

	(Control)	(Clique and Anticlique)	(Difference)
Enrolment	534.231	612.038	77.808
	(345.558)	(430.436)	(137.525)
Mean UCE '18-'20	0.064	0.098	0.034
	(0.636)	(0.415)	(0.169)
Private School	0.462	0.731	0.269
	(0.519)	(0.533)	(0.180)
No. Male Teachers	20.923	23.077	2.154
	(14.857)	(10.840)	(4.174)
No. Femaile Teachers	8.846	8.500	-0.346
	(4.776)	(5.736)	(1.849)
No. Teachers Eligible	20.692	20.692	0.000
	(14.285)	(12.985)	(4.559)
No. invited per year if treated	4.077	4.115	0.038
	(2.660)	(2.422)	(0.850)
Observations (schools)	13	26	39
RI p (Fisher's)			0.62

Notes: Standard deviations are reported in parentheses under group means and standard errors are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows.

Table 3.6. School Balance at Baseline: Anticlique vs Clique

	(Anticlique)	(Clique)	(Difference)
Enrolment	639.385	584.692	-54.692
	(459.384)	(416.336)	(171.950)
Mean UCE '18-'20	0.125	0.070	-0.054
	(0.369)	(0.470)	(0.166)
Private School	0.846	0.615	-0.231
	(0.555)	(0.506)	(0.208)
No. Male Teachers	23.154	23.000	-0.154
	(9.848)	(12.159)	(4.340)
No. Female Teachers	8.385	8.615	0.231
	(5.810)	(5.895)	(2.296)
No. Eligible Teachers	19.231	22.154	2.923
	(11.039)	(14.994)	(5.164)
No. invited per year if treated	3.846	4.385	0.538
	(2.075)	(2.785)	(0.963)
Observations (schools)	13	13	26
RI p (Fisher's)			0.52

Notes: Standard deviations are reported in parentheses under group means and standard errors are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows.

Table 3.7. Balance of baseline characteristics for students who took our 2022 student assessment.

	(Control)	(Clique or Anticlique)	(Difference)
Student Age	16.285	16.299	0.014
	(1.740)	(1.501)	(0.229)
Student Male	0.450	0.522	0.073
	(0.498)	(0.500)	(0.135)
S2	0.303	0.306	0.003
	(0.460)	(0.461)	(0.019)
S3	0.259	0.268	0.009
	(0.438)	(0.443)	(0.020)
S4	0.023	0.034	0.012
	(0.149)	(0.182)	(0.007)
School Mean UCE '18-'20	-0.192	0.117	0.309
	(0.753)	(0.379)	(0.303)
Observations (students)	3,774	7,721	11,495
Schools	13	25	38
RI p (Fisher's)			0.39

Notes: Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows.

Table 3.8. Balance for our sample of 2022 UCE Test-Takers

	(Control)	(Clique or Anticlique)	(Difference)
Student PLE Score	0.058	-0.031	-0.089
	(1.015)	(0.991)	(0.249)
Student Male	0.566	0.469	-0.097
	(0.496)	(0.499)	(0.105)
Student Age	18.995	19.003	0.008
	(1.432)	(1.376)	(0.188)
Schoom Mean UCE '18-'20	-0.065	0.176	0.241
	(0.690)	(0.382)	(0.245)
Observations	1,558	2,898	4,456
Schools	13	25	38
RI p (Fisher's)			.40

Notes: Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline school characteristics in the rows. Baseline UCE scores from '18-'20 contain imputations for two schools.

Table 3.9. Summary statistics for 2023 teacher survey

	mean	sd	min	max
School Mean UCE Score 2018-2020	-0.08	0.59	-1.43	0.87
Teacher Age	39.52	9.70	0.00	70.00
Teacher Male	0.67	0.46	0.00	1.00
Fraction agreeing with each statement:				
Boys have more ability than girls to learn science subjects	0.29	0.45	0.00	1.00
Boys need more education because they are better suited than girls to find high-paying jobs	0.18	0.39	0.00	1.00
Parents should maintain stricter control over their daughters than their sons	0.49	0.50	0.00	1.00
Girls should attain higher education so that they find better husbands	0.36	0.48	0.00	1.00
Boys should attain higher education so that they find better wives	0.27	0.44	0.00	1.00
Boys should be just as capable as girls at preparing a meal for visitors when guests visit	0.85	0.36	0.00	1.00
Girls have more ability than boys to learn science subjects	0.21	0.41	0.00	1.00
Girls need more education because they are better suited than boys to find high-paying jobs	0.26	0.44	0.00	1.00
Observations	455			

Notes: Mean UCE score was imputed as zero for two schools. Teacher age and gender were imputed at their means for nine teachers. No survey response was imputed.

Table 3.10. Balance of 2023 Teacher Survey by School Treatment Status

	(Pure Control)	(Clique or Anticlique)	(Difference)
Teacher Age	40.447	39.019	-1.427
	(8.409)	(10.303)	(1.640)
Teacher Male	0.623	0.703	0.080
	(0.486)	(0.450)	(0.063)
School Mean UCE '18-'20	-0.320	0.054	0.373
	(0.787)	(0.396)	(0.342)
Observations (teachers)	159	296	455
Schools	13	26	39
RI p (Fisher's)			0.30

Notes: Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a school was treated at all on the baseline characteristics in the rows

Table 3.11. Balance of 2023 Teacher Survey by Invitation Status

	(Not Invited)	(Invited)	(Difference)
Teacher Age	39.664	39.236	-0.428
	(9.285)	(10.476)	(1.317)
Teacher Male	0.649	0.725	0.076
	(0.475)	(0.440)	(0.054)
School Mean UCE '18-'20	-0.148	0.061	0.209
	(0.658)	(0.400)	(0.199)
Observations (teachers)	300	155	455
Schools	39	26	39
RI p (Fisher's)			0.55

Notes: Standard deviations are reported in parentheses under group means and standard errors clustered at the school level are reported under the differences. Fisher's exact p-value was calculated using the overall F-score of a regression of an indicator of whether a teacher had been invited in years 1 or 2 (2022 or 2023) on the baseline characteristics in the rows.

Table 3.12. Primary Hypothesis Tests for 2024.

Question	Test Stat	Interpretation	Sharp Null Hypothesis
Q1 a	$\hat{\Theta}_{Q1a.3.1}$	School assignment to treatment (Clique or Anticlique) did not affect any student exam scores on the '24 UCE exam.	$Y_{i3s}(\mathbf{T}) = Y_{i3s}(\mathbf{T}')$ if $\mathbf{1}\{D_s(\mathbf{T}) = \text{CTL}\} = \mathbf{1}\{D_s(\mathbf{T}') = \text{CTL}\}$ and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments
Q1 b	$\hat{\Theta}_{Q1b.3.2}$	Clique vs Anticlique school assignment did not affect any student exam scores on the '24 UCE exam.	$Y_{i3s}(\mathbf{T}) = Y_{i3s}(\mathbf{T}')$ if $D_s(\mathbf{T}) \neq$ CTL and $D_s(\mathbf{T}') \neq$ CTL and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments
Q1 c	$\hat{\Theta}_{Q1c.3.3}$	School assignment to treatment did not affect any girls' exam scores on the '24 UCE exam.	$Y_{i3s}(\mathbf{T}) = Y_{i3s}(\mathbf{T}')$ if $\mathbf{1}\{D_s(\mathbf{T}) = \text{CTL}\} = \mathbf{1}\{D_s(\mathbf{T}') = \text{CTL}\}$ and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments for girls
Q2 a	$F_{Q2a.3.8}$	An invitation to training did not affect any teacher survey responses to gender-related questions in '24.	$Z_{i3su}(\mathbf{T}) = Z_{i3su}(\mathbf{T}')$ for all $\mathbf{T}, \mathbf{T}', u$
Q2 b	$F_{Q2b.3.9}$	Clique vs Anticlique assignment did not affect any survey responses to gender-related questions among invited teachers in '24.	$Z_{i3su}(\mathbf{T}) = Z_{i3su}(\mathbf{T}')$ if $D_s(\mathbf{T}) \neq$ CTL and $D_s(\mathbf{T}') \neq$ CTL and $\mathbf{T}, \mathbf{T}'$ have same teacher cluster assignments
Q2 c	$F_{T.Q2c.3.10}$	Teacher training assignments did not affect teacher survey responses to any gender-related questions for uninvited teachers in '24.	$Z_{i3su}(\mathbf{T}) = Z_{i3su}(\mathbf{T}')$ if $T_{j3} = T'_{j3} = 0$

Notes: Every hypothesis will be tested using Fisher's Exact Test using the test statistic in the middle column. Additional followup exploratory tests will be run, but these tests are specified in advance so they are non-exploratory.

Table 3.13. Midline results and endline power calculations for Q1 a: main effects on student exam scores.

ITT effect of treatment school vs control school on 2022 exams

		Midline Estimate				Power Calc	
Estimand	Exam	N	Estimate	(se)	RI p	Effect	Pwr
$\Theta_{Q1a.3.1}$	Ours	11495	.096	(.119)	.74	.32	92%
$\Theta_{Q1a.3.1}$	UCE	4456	.014	(.067)	.89	.24	93%

Notes: Compares exams scores of students in treatment (clique or anticlique) schools to students in pure control schools. The outcome is the student exam score expressed in standard deviations from the sample mean. Standard errors are clustered at the school level and reported in parentheses. The left panel shows the midline estimate. The ITT was estimated by regression-adjustment. The top row uses our 2022 assessment as the outcome and the bottom row uses the 2022 official UCE exam scores. When the outcome variable is our assessment, covariates included school mean baseline UCE scores and student age, gender and cohort. When the outcome variable is the UCE exam, covariates include school mean baseline UCE scores, student PLE score, and student age and gender. The right panel shows a conservative power calculation based on resampling the 2022 data with replacement within schools and adding an additional effect to treated schools.

Table 3.14. Midline results and endline power calculations for Q1 b: effect of Clique training on student exam scores.

		Midline Estimate				Power Calc	
Estimand	Exam	N	Estimate	(se)	RI p	Effect	Pwr
$\Theta_{Q1b.3.2}$	Ours	7721	.190	(.158)	.49	.38	90%
$\Theta_{Q1b.3.2}$	UCE	2898	.109	(.082)	.53	.19	95%

Notes: Compares scores of students in Clique schools to students in Anticlique schools. The outcome is the student exam score expressed in standard deviations from the sample mean. Standard errors are clustered at the school level and reported in parentheses. The left panel shows the midline estimate. The ITT was estimated by regression-adjustment. The top row uses our 2022 assessment as the outcome and the bottom row uses the 2022 official UCE exam scores. When the outcome variable is our assessment, covariates included school mean baseline UCE scores and student age, gender and cohort. When the outcome variable is the UCE exam, covariates include school mean baseline UCE scores, student PLE score, and student age and gender. The right panel shows a conservative power calculation based on resampling the 2022 data with replacement within schools and adding an additional effect to treated schools.

Table 3.15. Midline results and endline power calculations for Q1 c: main effects on girls' exam scores.

		Midline Estimate				Power Calc	
Estimand	Exam	N	Estimate	(se)	RI p	Effect	Pwr
$\Theta_{Q1c.3.3}$	Ours	5765	-.016	(.067)	.84	.25	93%
$\Theta_{Q1c.3.3}$	UCE	2214	.031	(.100)	.94	.39	95%

Notes: Compares exams scores of girls in treatment (clique or anticlique) schools to girls in pure control schools. The outcome is the exam score expressed in standard deviations from the sample mean. Standard errors are clustered at the school level and reported in parentheses. The left panel shows the midline estimate. The ITT was estimated by regression-adjustment. The top row uses our 2022 assessment as the outcome and the bottom row uses the 2022 official UCE exam scores. When the outcome variable is our assessment, covariates included school mean baseline UCE scores and student age and cohort. When the outcome variable is the UCE exam, covariates include school mean baseline UCE scores, student PLE score, and student age. The right panel shows a conservative power calculation based on resampling the 2022 data with replacement within schools and adding an additional effect to treated schools.

Table 3.16. Midline results for Q2 a: main effects on teacher gender attitudes.

Estimand	Survey question	Estimate	(s.e.)	RI p
$\Theta_{Q2a.3.4.u}$	Boys have more ability than girls to learn science subjects	.019	(.080)	.80
$\Theta_{Q2a.3.4.u}$	Boys need more education because they are better suited than girls to find high-paying jobs	-.042	(.076)	.57
$\Theta_{Q2a.3.4.u}$	Parents should maintain stricter control over their daughters than their sons	-.032	(.04)	.52
$\Theta_{Q2a.3.4.u}$	Girls should attain higher education so that they find better husbands	.022	(.053)	.82
$\Theta_{Q2a.3.4.u}$	Boys should attain higher education so that they find better wives	-.035	(.052)	.51
$\Theta_{Q2a.3.4.u}$	Boys should be just as capable as girls at preparing a meal for visitors when guests visit	.021	(.047)	.68
$\Theta_{Q2a.3.4.u}$	Girls have more ability than boys to learn science subjects	.131	(.073)	.10
$\Theta_{Q2a.3.4.u}$	Girls need more education because they are better suited than boys to find high-paying jobs	-.101	(.097)	.20
$F_{Q2a.3.8}$	Overall test of H_0: no effect on any teacher or question			.01**
N	Teachers (excluding non-invitees in treated schools)			314

Notes: Compares invited teachers vs teachers in pure control schools. Outcomes are the fraction of teachers agreeing with different gender-related statements. Covariates were teacher age, teacher gender, and school mean baseline UCE score. Standard errors clustered at the school level in parentheses. Data: midline 2023 teacher survey.

Table 3.17. Midline results for Q2 b: Clique effects on teacher gender attitudes.

Estimand	Survey question	Estimate	(s.e.)	RI p
$\Theta_{Q2b.3.5.u}$	Boys have more ability than girls to learn science subjects	-.186	(.071)	.11
$\Theta_{Q2b.3.5.u}$	Boys need more education because they are better suited than girls to find high-paying jobs	-.255	(.068)	.00***
$\Theta_{Q2b.3.5.u}$	Parents should maintain stricter control over their daughters than their sons	-.121	(.071)	.10*
$\Theta_{Q2b.3.5.u}$	Girls should attain higher education so that they find better husbands	-.278	(.076)	.00***
$\Theta_{Q2b.3.5.u}$	Boys should attain higher education so that they find better wives	-.084	(.065)	.28
$\Theta_{Q2b.3.5.u}$	Boys should be just as capable as girls at preparing a meal for visitors when guests come to visit	.026	(.059)	.67
$\Theta_{Q2b.3.5.u}$	Girls have more ability than boys to learn science subjects	-.127	(.056)	.19
$\Theta_{Q2b.3.5.u}$	Girls need more education because they are better suited than boys to find high-paying jobs	-.209	(.071)	.04**
$F_{Q2b.3.9}$	Overall test of H_0: no Clique effect on any teacher or question			.02**
N	Teachers (excluding pure control schools)			155

Notes: Compares invited teachers in Clique vs Anticlique schools. Outcomes are the fraction of teachers agreeing with different gender-related statements. Standard errors clustered at the school level in parentheses. Covariates included school mean baseline UCE score and teacher age and gender. Data: 2023 teacher survey.

Table 3.18. Midline results for Q2 c: spillovers on teacher gender attitudes.

Estimand	Survey question	Estimate	(s.e.)	RI p
$\Theta_{Q2c.3.6.u}$	Boys have more ability than girls to learn science subjects	.065	(.075)	.41
$\Theta_{Q2c.3.6.u}$	Boys need more education because they are better suited than girls to find high-paying jobs	-.074	(.055)	.28
$\Theta_{Q2c.3.6.u}$	Parents should maintain stricter control over their daughters than their sons	-.009	(.041)	.85
$\Theta_{Q2c.3.6.u}$	Girls should attain higher education so that they find better husbands	.115	(.081)	.16
$\Theta_{Q2c.3.6.u}$	Boys should attain higher education so that they find better wives	-.052	(.061)	.30
$\Theta_{Q2c.3.6.u}$	Boys should be just as capable as girls at preparing a meal for visitors when guests visit	.061	(.044)	.20
$\Theta_{Q2c.3.6.u}$	Girls have more ability than boys to learn science subjects	.092	(.054)	.20
$\Theta_{Q2c.3.6.u}$	Girls need more education because they are better suited than boys to find high-paying jobs	-.1	(.057)	.21
$F_{Q2c.3.10}$	Overall test of H_0: no spillover effect on any uninvited teacher for any question			.12
N	Teachers (excluding invitees)			300

Notes: Compares uninvited teachers in treatment vs control schools. Outcomes are the fraction of teachers agreeing with gender-related statements in second column. Standard errors clustered at the school level in parentheses. Data: 2023 teacher survey.

Chapter 3, in full, is an unpublished report issued midway through a multi-year field experiment. The dissertation author wrote the full text of this chapter and provided technical work at many stages of the field experiment from October 2021 until the present. The co-authors are Moustafa El-Kashlan, Vesall Nourani, and Sara Restrepo-Tamayo.

Bibliography

Alan, Sule and Ipek Mumcu, "Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention," *American Economic Review*, April 2024, *114* (4), 1173–1210.

Allen, Treb, Costas Arkolakis, and Yuta Takahashi, "Universal Gravity," *Journal of Political Economy*, 2020, *128* (2), 393–433.

Andrews, Isaiah and Maximilian Kasey, "Identification of and Correction for Publication Bias," *American Economic Review*, August 2019, *109* (8), 2766–94.

_ **and Maximilian Kasy**, "Identification of and Correction for Publication Bias," *American Economic Review*, August 2019, *109* (8), 2766–94.

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, Marco Mendoza Aviña, and Tom D. Stanley, "Quantitative Political Science Research is Greatly Underpowered," I4R Discussion Paper Series 6, The Institute for Replication (I4R) 2022.

Armstrong, Timothy B. and Michal Kolesár, "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*, 2020, *11* (1), 1–39.

Aronow, Peter M. and Cyrus Samii, "Estimating average causal effects under general interference, with application to a social network experiment," *The Annals of Applied Statistics*, 2017, *11* (4), 1912 – 1947.

Baird, Sarah, Joan Hamory Hicks, Michael Kremer, and Edward Miguel, "Worms at Work: Long-run Impacts of a Child Health Investment*," *The Quarterly Journal of Economics*, 07 2016, *131* (4), 1637–1680.

Basse, Guillaume W. and Edoardo M. Airoldi, "Limitations of Design-based Causal Inference and A/B Testing under Arbitrary and Network Interference," *Sociological Methodology*, 2018, *48* (1), 136–151.

Bateman, Harry, *Higher Transcendental Functions [Volumes I-III]*, McGraw-Hill Book Company, 1953. Funding by Office of Naval Research (ONR).

Borusyak, Kirill and Peter Hull, "Nonrandom Exposure to Exogenous Shocks," *Econometrica*, 2023, *91* (6), 2155–2185.

Bramoullé, Yann, Rachel Kranton, and Martin D'Amours, "Strategic Interaction and Networks," *American Economic Review*, March 2014, *104* (3), 898–930.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, January 2016, *8* (1), 1–32.

_ , Nikolai Cook, and Anthony Heyes, "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, November 2020, *110* (11), 3634–60.

Brunner, Jerry and Ulrich Schimmack, "Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance," *Meta-Psychology*, 05 2020, *4*.

Businge, Conan, "What is the New National Teachers' Policy," *The New Vision*, 2019.

Carrasco, Marine and Jean-Pierre Florens, "A spectral method for deconvolving a density," *Econometric Theory*, 2011, *27* (3), 546–581.

_ , _ , and Eric Renault, "Chapter 77 Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in James J. Heckman and Edward E. Leamer, eds., *James J. Heckman and Edward E. Leamer, eds.*, Vol. 6 of *Handbook of Econometrics*, Elsevier, 2007, pp. 5633–5751.

de Chaisemartin, Clément and Xavier D'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, September 2020, *110* (9), 2964–96.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia*," *The Quarterly Journal of Economics*, 11 2017, *133* (2), 993–1039.

DellaVigna, Stefano and Elizabeth Linos, "RCTs to Scale: Comprehensive Evidence From Two Nudge Units," *Econometrica*, 2022, *90* (1), 81–116.

Donaldson, Dave and Richard Hornbeck, " Railroads and American Economic Growth: A "Market Access" Approach *," *The Quarterly Journal of Economics*, 02 2016, *131* (2), 799–858.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael

Walker, "General Equilibrium Effects of Cash Transfers: Experimental Evidence From Kenya," *Econometrica*, 2022, *90* (6), 2603–2643.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich, "Detecting p-Hacking," *Econometrica*, 2022, *90* (2), 887–906.

Fan, Jianqing, "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *The Annals of Statistics*, 1991, *19* (3), 1257 – 1272.

Ferraro, Paul J. and Pallavi Shukla, "Credibility crisis in agricultural economics," *Applied Economic Perspectives and Policy*, 2023, *45* (3), 1275–1291.

Florens, Jean-Pierre, Joël L. Horowitz, and Ingrid Van Keilegom, "Bias-Corrected Confidence Intervals in a Class of Linear Inverse Problems," *Annals of Economics and Statistics*, 2017, (128), 203–228.

Franco, Annie, Neil Malhotra, and Gabor Simonovits, "Publication bias in the social sciences: Unlocking the file drawer," *Science*, 2014, *345* (6203), 1502–1505.

Gerber, Alan and Neil Malhotra, "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals," *Quarterly Journal of Political Science*, 2008, *3* (3), 313–326.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin, "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, January 2009, *1* (1), 112–35.

Goldsmith-Pinkham, Paul and Guido Imbens, "Social Networks and the Identification of Peer Effects," *Journal of Business, Economics and Statistics*, 2013.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277. Themed Issue: Treatment Effect 1.

Gupta, Sabhya and Sarah Kopper, https://www.povertyactionlab.org/resource/power-calculations 2023. Accessed: 20123-04-30.

Hajek, Jároslav, "Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One."," in V. P. Godambe and D. A. Sprott, eds., *Foundations of Statistical Inference*, Holt, Rinehart and Winston, 1971.

Hoenig, John M and Dennis M Heisey, "The Abuse of Power," *The American Statistician*, 2001, *55* (1), 19–24.

Indritz, Jack, "An Inequality for Hermite Polynomials," *Proceedings of the American Mathematical Society*, 1961, *12* (6), 981–983.

Ioannidies, John P. A., T. D. Stanley, and Hristos Doucouliagos, "The Power of Bias in Economics Research," *The Economic Journal*, October 2017, *127*, F236–F265.

Jagadeesan, Ravi, Natesh S. Pillai, and Alexander Volfovsky, "Designs for estimating the treatment effect in networks with interference," *The Annals of Statistics*, 2020, *48* (2), 679 – 712.

Jenish, Nazgul and Ingmar R. Prucha, "On spatial processes and asymptotic inference under near-epoch dependence," *Journal of Econometrics*, 2012, *170* (1), 178–190.

Johnston, William, "The Weighted Hermite Polynomials Form a Basis for L2(R)," *The American Mathematical Monthly*, 2014, *121* (3), pp. 249–253.

Kelejian, Harry and Gianfranco Piras, "Chapter 1 - Spatial Models: Basic Issues**Basic texts in spatial analysis and econometrics are Cliff and Ord, 1973, Cliff and Ord, 1981, Anselin (1988), and Cressie (1993). More recent texts and compilations are Anselin and Florax (1995b), Anselin et al. (2004), Anselin and Rey (2014), LeSage and Pace, 2004, LeSage and Pace, 2009, Elhorst (2014), and Arbia, 2006, Arbia, 2014. See also a nice overview of spatial models, and the development of spatial econometrics by Anselin, 2002, Anselin, 2009 and Anselin and Bera (1998).," in Harry Kelejian and Gianfranco Piras, eds., *Spatial Econometrics*, Academic Press, 2017, pp. 1–10.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, A. E. van 't Veer, Leigh Ann Vaughn, Marek Vranka, Aaron L. Wichman, Julie A. Woodzicka, and Brian A. Nosek, "Investigating Variation in Replicability," *Social Psychology*, 2014, *45* (3), 142–152.

Kremer, Michael, Conner Brannen, and Rachel Glennerster, "The Challenge of Education and Learning in the Developing World," *Science*, 2013, *340* (6130), 297–300.

Kudrin, Nikolay, "Robust Caliper Tests," *Working Paper*, 2023.

Lang, Kevin, "How Credible is the Credibility Revolution?," Working Paper 31666, National Bureau of Economic Research September 2023.

Lee, Lung-fei, "GMM and 2SLS estimation of mixed regressive, spatial autoregressive models," *Journal of Econometrics*, 2007, *137* (2), 489–514.

Leung, Michael P., "Causal Inference Under Approximate Neighborhood Interference," *Econometrica*, 2022, *90* (1), 267–293.

_ , "Rate-Optimal Cluster-Randomized Designs for Spatial Interference," 2022.

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

_ , "Identification of treatment response with social interactions," *The Econometrics Journal*, 2013, *16* (1), S1–S23.

Manta, Alexandra, Anson T.Y. Ho, Kim P. Huynh, and David T. Jacho-Chávez, "Estimating social effects in a multilayered Linear-in-Means model with network data," *Statistics and Probability Letters*, 2022, *183*, 109331.

McEwan, Patrick J., "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments," *Review of Educational Research*, 2015, *85* (3), 353–394.

Miguel, Edward and Michael Kremer, "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities," *Econometrica*, 2004, *72* (1), 159–217.

Muralidharan, Karthik and Paul Niehaus, "Experimentation at Scale," *Journal of Economic Perspectives*, November 2017, *31* (4), 103–24.

_ , _ , and Sandip Sukhtankar, "General Equilibrium Effects of (Improving) Public Employment Programs: Experimental Evidence from India," Working Paper 23838, National Bureau of Economic Research September 2021.

Najjumba, Innocent Mulindwa and Jeffery H. Marshall, "Improving Learning in Uganda Vol. II: Problematic Curriculum Areas and Teacher Effectiveness: Insights from National Assessments," Technical Report, World Bank 2013.

Neumann, Michael H., "A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics," *ESAIM: Probability and Statistics*, 2013, *17*, 120–134.

Newey, Whitney K. and Daniel McFadden, "Chapter 36 Large sample estimation and hypothesis testing," in "in," Vol. 4 of *Handbook of Econometrics*, Elsevier, 1994, pp. 2111–2245.

Nourani, Vesall, Nava Ashraf, and Abhijit Banerjee, "Learning to Teach by Learning to Learn," *Working Paper*, 2023.

Popova, Anna, David K Evans, Mary E Breeding, and Violeta Arancibia, "Teacher Professional Development around the World: The Gap between Evidence and Practice," *The World Bank Research Observer*, 06 2021, *37* (1), 107–136.

Racine, Jeffrey S., Liangjun Su, and Aman Ullah, *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*, Oxford University Press, 02 2014.

Rooke, Barnaby, "Teacher Issues in Uganda: A shared vision for an effective teachers policy," Technical Report, Ugandan Ministry of Education and Sport, UNESCO 2014.

Sotola, Lukas, "How Can I Study from Below, that which Is Above?: Comparing Replicability Estimated by Z-Curve to Real Large-Scale Replication Attempts," *Meta-Psychology*, 09 2023, *7*.

Sussman, Daniel L. and Edoardo M. Airoldi, "Elements of estimation theory for causal effects in the presence of network interference," 2017.

Sävje, Fredrik, "Causal inference with misspecified exposure mappings," 2023.

Vasiliev, Kirill and Angela Demas, "Uganda Secondary Education Expansion Project (P166570)," Technical Report, The World Bank 2018.

Wand, Matt P. and M. Chris Jones, *Kernel Smoothing* 1995.

Wooldridge, Jefrey M., *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, 2010.

World Bank, "Lower secondary completion rate. Series SE.SEC.CMPT.LO.ZS," 2024.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results*," *The Quarterly Journal of Economics*, 11 2018, *134* (2), 557–598.

Zhang, Lina, "Spillovers of Program Benefits with Mismeasured Networks," 2023.