Fall 10-1-2021

Algorithmic Advances for the Design and Analysis of Randomized Experiments

Christopher Robert Harshaw
Yale University Graduate School of Arts and Sciences, crharshaw@gmail.com

Abstract
Algorithmic Advances for the
Design and Analysis of Randomized Experiments
Christopher Harshaw
2021

Randomized experiments are the gold standard for investigating the causal effect of treatment on a population. In this dissertation, we present algorithmic advances for three different problems arising in the design and analysis of randomized experiments: covariate balancing, variance estimation, and bipartite experiments.

In the first chapter, we describe an inherent trade-off between covariate balancing and robustness, which we formulate as a distributional discrepancy problem. In order to navigate this trade-off, we present the Gram–Schmidt Walk Design which is based on the recent discrepancy algorithm of Bansal, Dadush, Garg, and Lovett (2019). By tightening the algorithmic analysis, we derive bounds on the mean squared error of the Horvitz–Thompson estimator under this design in terms of a ridge regression of the outcomes on the covariates, which we interpret as regression by design. We carry out further analysis including tail bounds on effect estimator, methods for constructing confidence intervals, and an extension of the design which accommodates non-linear responses via kernel methods.

In the second chapter, we study the problem of estimating the variance of treatment effect estimators under interference. It is well-known that unbiased variance estimation is impossible without strong assumptions on the outcomes, due to the fundamental problem of causal inference. Thus, we study a class of conservative estimators which are based on variance bounds. We identify conditions under which

the variance bounds themselves are admissible and provide a general algorithmic framework to construct admissible variance bounds, according to the experimenter's preferences and prior substantive knowledge.

In the final chapter, we present methodology for the newly proposed bipartite experimental framework, where units which receive treatment are distinct from units on which outcomes are measured, and the two are connected via a bipartite graph. We investigate a linear exposure-response assumption which allows more complex interactions. We propose the Exposure Re-weighted Linear (ERL) estimator which we show is unbiased in finite samples and consistent and asymptotically normal in large samples provided the bipartite graph is sufficiently sparse. We provide a variance estimator which facilitates confidence intervals based on the normal approximation. Finally, we present Exposure-Design, a correlation clustering based design for improving precision of the ERL estimator.

Algorithmic Advances for the Design and Analysis of Randomized Experiments

A Dissertation
Presented to the Faculty of the Graduate School
Of
Yale University
In Candidacy for the Degree of
Doctor of Philosophy

By
Christopher Harshaw

Dissertation Directors: Daniel Spielman and Amin Karbasi

December 2021

In loving memory of my grandfather, José Garcia Bacallao, who loved to tell stories, to laugh, to sing, and to sit on the beach.

Acknowledgment

First and foremost, I am eternally grateful to my academic advisors, Daniel Spielman and Amin Karbasi. Among my fondest memories during my time at Yale will be the countless hours we spent throwing conjectures on the whiteboard with an open Jupyter notebook terminal. Truly, I am a better researcher and a better person to have known and worked so closely with you both. Should I be lucky enough to advise students one day, I hope to pass on the virtues of intellectual curiosity, academic rigor, and kindness towards others that you have shown to me. I would also like to extend my sincerest gratitude towards the other members of my dissertation committee, Fredrik Sävje and Sekhar Tatikonda. My collaborations with Fredrik, which culminated in this dissertation, have been some of my most exciting and meaningful work.

Thank you to my many collaborators from whom I have learned so much. In particular, I want to thank Peng Zhang, Joel Middleton, Jean Pouget-Abadie, Sébastien Lahaie, Vahab Mirrokni, and David Eisenstat for countless hours at the whiteboard, insightful discussions, and mentorship. I also thank Moran Feldman, Lin Chen, Hamed Hassani, and Justin Ward for the collaborations on submodular optimization, which did not appear in this dissertation but were a joy to work on. I thank Rasmus Kyng for showing me the ropes when I was a fresh and confused PhD student. I also thank Tim Kunisky for carefully proof reading a first draft of this dissertation and catching many errors. A special thanks goes to Ehsan Kazemi and Marko Mitrovic, who have been the best conference travel buddies.

Thank you to my mother and father for their continued love and support during these past six years. None of this would be possible without your hard work, both during my childhood and to this day. You two have given me every opportunity, and words cannot begin to express my gratitude. I thank Meghan for her love and support as my sister and my oldest friend. I also thank my extended family—my grandmothers, uncles, aunts, and cousins—for cheering me on the past six years.

I extend my warmest thanks to my friends, whom I dearly cherish. I thank the LoFi boys Alex, Henry, and Jake for the paradigm-shifting Mark Fisher-fueled discussions, one of the coolest road trips ever, and the most sincere comradery I could have asked for. I thank the KBT Crew for all the nights spent drinking beer by the fire, typically discussing anime and extraterrestrial life. I thank members of The Clinic for the memorable jam sessions and for time spent in the pocket. Thank you to Josie Bircher, Joe Sullivan, and Chloe Larkin for their companionship as the best

roommates in New Haven. A special shout out goes to the Jitter Bus boys Dan, Paul, AJ, and Andrew as well as the extended Jitter Bus crew for the caffeine, friendship, and mosh pits. Finally, I thank Jon Erickson for his friendship during the past nearly two decades and for keeping me mostly sane during the early isolated stages of the pandemic by playing video games with me late into the night.

I extend my sincerest thanks to those who have helped me grow in fundamental ways. Thank you to Alida Engel, Gene Burger, Danette Fitzgerald and folks in the New Haven chapter of the National Stuttering Association for helping me to find my voice and to *celebrate* it. Thank you to the gains squad Lyle Huneke, AJ Tarantino, Brittany Hutchison, and Zac Riegelmann for teaching me how to tap into my inner strength and achieve new goals.

Last but most certainly not least, I would like to thank Julia Stevenson for her undying love and support during these years—her kindness, compassion, and joy have made my world much brighter. I extend my gratitude to the entire Stevenson family for welcoming me as one of their own.

Funding I gratefully acknowledge the funding that supported the research which appears in this dissertation. In particular, I acknowledge NSF Graduate Research Fellowship (DGE1122492) awarded to me as well as support from Google as a Summer Intern and a Student Researcher. In addition, I acknowledge NSF Grant CCF-1562041, ONR Awards N00014-16-2374 and N00014-20-1-2335, and a Simons Investigator Award awarded to my generous advisor, Daniel Spielman.

Contents

Chapter 1

Introduction

1.1 Overview

Randomized experiments are widely regarded as the gold standard for investigating the causal effect of a treatment on a population. Since their modern inception in the early twentieth century, randomized experiments have been used in a wide variety of fields, from agricultural and medical research to the economics of global development.

Many, if not most, statistical problems in the design and analysis of randomized experiments are inherently computational in nature. One of the primary problems is to construct an *experimental design*—that is, the distribution of random treatment assignments (say, drug or placebo) to participants in the experiment—to meet certain specifications. There are a number of possible specifications an experimenter may desire: one is to randomly assign participants to treatment and control groups which are, with high probability, similar with respect to all observable characteristics. Another specification is to limit the interaction between the treatment groups, given a fixed interaction model. When only two treatment assignments are considered in the experiment, these problems may be rephrased as constructing a distribution over vertices of the hypercube which meet certain specifications. How to construct an efficient sampling algorithm meeting these specifications, or even to determine whether such a distribution exists, are computational questions which may be informed by the growing body of work on sampling algorithms in theoretical computer science.

In this dissertation, I focus on three statistical problems arising in the design and analysis of randomized experiments, which are fundamentally computational in nature:

- **Covariate Balancing**: Experimenters can often obtain pre-treatment covariates of the subjects, e.g. age, weight, income, gender. Under a fully randomized experimental design, randomly assigned treatment groups typically differ in one or more of the covariates. Experimenters sometimes choose experimental designs which are less random, but produce more balanced treatment groups. *How does balancing covariates between treatment groups affect precision of the treatment effect estimator? Can we construct an experimental design which balances*

covariates in a near-optimal manner? In Chapter 2, we describe an inherent trade-off between covariate balance and robustness in experimental design and present GSW-DESIGN, a discrepancy-theoretic experimental design which provably navigates this trade-off in a near-optimal way.

- **Variance Estimation**: The treatment effect is estimated via a randomized experiment, so our estimator of the effect is a random variable. In order to investigate the precision of the estimator and construct confidence intervals, an experimenter needs to have a good sense of its variance. *How can we estimate the variance of our estimator from just one run of the experiment?* This problem becomes more challenging in the presence of interference, where interactions between units affects observed outcomes. In Chapter 3, we show that no universally best variance estimator exists and provide an optimization framework for constructing an admissible variance estimator under arbitrary interference.

- **Bipartite Experiments**: The recently introduced bipartite experimental framework formalizes experimental settings where the units of treatment differ from the units on which outcomes are measured. This scenario arises in a variety of research areas from public health interventions to marketplace strategies. *How can we estimate treatment effects under complex interactions between the two groups of units?* In Chapter 4, we present the Exposure Reweighted Linear (ERL) estimator and EXPOSURE-DESIGN, the later based on a correlation-clustering formulation, for estimating treatment effects in a bipartite experiment under a linear exposure-response assumption.

The connections made between experimental design and computation in this dissertation only scratch the surface. Experimental methods have a wealth of problems that can be informed by the use of sophisticated computational techniques. Likewise, these statistical problems may drive new developments in the study of algorithms. It is my opinion that this connection will provide many fruitful research directions.

The intended audience of this dissertation are statistical methodologists working in design-based causal inference as well as computer scientists interested in algorithm design and analysis. It is my hope that the causal inference problems we consider and the computational techniques we propose will be relevant and interesting to both communities. For the sake of clarity and conciseness, I have chosen to focus on the presentation and interpretation of mathematical results, rather than the techniques used to prove them. The majority of proofs appear in the appendix, although shorter, more instructive proofs appear in the main body.

The remainder of this chapter contains an introductory treatment of the potential outcomes framework. This will be useful to the uninitiated, but may be skipped by experts. The remaining three chapters are largely independent and self contained: each chapter begins with an introduction of relevant preliminary material and ends with a concluding discussion of open problems.

1.2 Preliminaries

1.2.1 Potential outcomes framework

In this dissertation, I focus on the potential outcomes framework for causal inference, first formulated by Neyman (1923) and later independently discovered and pioneered by Rubin (1974). For a more complete treatment of randomized experiments under the potential outcome framework, we refer readers to Imbens and Rubin (2015).

Before formally defining the framework, I would like to take some time to discuss the types of causal questions which experimenters seek to investigate when invoking this framework. Holland (1986) describes two types of causal reasoning: discovering the cause of a given effect and determining the effect of a given cause. Generally speaking, the former is much more challenging and should seem to require far more assumptions. Let me further describe both of these types of causal reasoning by way of example: imagine a team of doctors is responsible for and closely monitors the health of several hundred cancer patients. Upon seeing a significant improvement in the rates of cancer remission among the patients—measured by the circulating tumor cell (CTC) count—the team of doctors might ask the causal question: "what caused the CTC count to decrease?" This is a very challenging question because it requires, among other things, the doctors' knowledge of every possible variable involved in the presumed underlying causal mechanism of the patient's cancer. On the other hand, the doctors may ask a more targeted causal question: "what is the effect of a six month experimental drug course on the CTC count?" This is a more tractable question because here the cause is fixed and the effect is to be determined. One approach to answering this question is to design a study which explicitly controls this cause by assigning (typically at random) which patients receive the experimental drug. The potential outcomes framework addresses these styles of targeted causal questions, which are the focus of this dissertation.

The potential outcome framework is now formally defined: there are n experimental units which are index by integers $i \in \{1, 2, \dots n\} = [n]$. In our example above, the experimental units are the cancer patients. Each unit receives a binary treatment $z_i \in \{\pm 1\}$ and we collect the n treatments into a treatment vector,

$$\boldsymbol{z} = (z_1, z_2, \dots z_n) \ .$$

The treatment z_i assigned to unit $i \in [n]$ is random, and so the treatment vector $\boldsymbol{z}$ is a random vector. Let $Z^+ = \{i \in [n] : z_i = 1\}$ and $Z^- = \{i \in [n] : z_j = -1\}$ be the random partition of the units into the treated and control groups. The two treatments in our example are a six month course of the experimental drug ($z_i = 1$) and a six month course of a placebo ($z_i = -1$). It is important to note that, the experiment is only run once, meaning that only one treatment vector $\boldsymbol{z} \in \{\pm 1\}^n$ is selected. The distribution over treatment vectors is specified by the experimenter and is referred to as the *experimental design*, or simply the *design*.

Every unit has two *potential outcomes* which are associated with the two treatments. If unit $i \in [n]$ is assigned treatment $z_i = 1$, then the experimenter observes potential outcome $a_i \in \mathbb{R}$; otherwise, the unit is assigned $z_i = -1$ and the potential outcome $b_i \in \mathbb{R}$ is observed. The two outcomes are referred to as "potential" because either could be potentially observed, but only one is actually observed after treatment is assigned. In our example, the two outcomes for each unit are the CTC counts of the patients under the drug and the placebo treatments. For each unit $i \in [n]$, the observed outcome is a random variable which depends on the treatment,

$$y_i = \begin{cases} a_i & \text{if } z_i = 1 \\ b_i & \text{if } z_i = -1 \end{cases} .$$

We find it useful to collect the the two (deterministic) potential outcomes and the (random) observed outcome into vectors:

$$\boldsymbol{a} = (a_1, a_2, \ldots a_n) \quad \text{and} \quad \boldsymbol{b} = (b_1, b_2, \ldots b_n) \quad \text{and} \quad \boldsymbol{y} = (y_1, y_2, \ldots y_n) \ .$$

We also assume that for each unit $i \in [n]$, the experimenter has collected a d-dimensional covariate vector $\boldsymbol{x}_i \in \mathbb{R}^d$. The covariate vectors are known before treatment and so they may be used in the construction of both the design and the estimator. In our example, the covariates may include age, gender, and weight of the participants in addition to results of pre-treatment diagnostic tests.

We emphasize here that the units, their outcomes, and their covariates are fixed, deterministic quantities. All randomness is induced by the random assignment of treatment by the experimenter. Additionally, the potential outcomes framework, as we have described it here, has implicity invoked the Stable Unit Treatment Value Assumption (SUTVA) which dictates that the potential outcomes are well-defined (i.e. there is no hidden or unobserved treatment) and also that the outcome of a unit is determined solely by its own treatment assignment (Rubin, 1980; Holland, 1986). Experiments which violate this second part of the SUTVA are said to exhibit *interference*. The design and analysis of experiments under interference are considered in Chapters 3 and 4.

We now define several causal quantities of interest in the potential outcome framework. The first such quantity is the *individual treatment effect* (ITE), which is defined as the contrast between a unit's treatment and control outcomes,

$$\tau_i = a_i - b_i \quad \text{for all } i \in [n] \ .$$

A unit's individual treatment effect is never directly observed because only one of these potential outcomes is observed. Moreover, we cannot hope to estimate each unit's ITE within reasonable precision. In our example, a patient's ITE is the difference in the CTC counts under the drug course and the placebo course. An aggregate

causal quantity is the *average treatment effect* (ATE), which is defined as

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i) \ .$$

The ATE cannot be observed directly because only half of the outcomes are observed; however, by constructing an appropriate experimental design, the experimenter can hope to estimate this quantity to a reasonable precision. In the cancer trial example, the ATE is the effect of the experimental drug course on CTC counts, averaged over all patients in the study.

The overall goal is to construct a design and an estimator so that the ATE is estimated to reasonably high precision. There are a myriad of designs and estimators which experimenters use for this task. For concreteness, we discuss a few of the most common choices.

Designs There are a number of designs which experimenters use in practice, but the two most common designs are the Bernoulli design and the group-balanced design.

- **Bernoulli Design**: In the Bernoulli design, units are assigned treatments uniformly and independently. In this way, the probability of each assignment vector $\boldsymbol{z} \in \{\pm 1\}^n$ under the Bernoulli design is $1/2^n$.
- **Group-Balanced Design**: In the group-balanced design, a uniformly random half of the units are selected to receive treatment. Thus, the treatment group Z^+ follows the distribution $Z^+ \sim \text{Unif}\{Z \subset \{\pm 1\}^n : |Z| = n/2\}$. It is typically assumed that when using a group-balanced design, n is even.

Estimators A wide variety of estimators are used in practice. Two of the most common estimators are the Horvitz–Thompson and difference-in-means estimators.

- **Horvitz–Thompson**: The Horvitz–Thompson estimator is a difference between the observed outcomes in the treatment and control groups, weighted by the inverse of the probability of observing that treatment. More formally,

$$\widehat{\tau} = \frac{1}{n}\left[\sum_{i\in Z^+}\frac{a_i}{\Pr(z_i = 1)} - \sum_{i\in Z^-}\frac{b_i}{\Pr(z_i = -1)}\right] \ .$$

  As we will show later in this section, the re-weighting terms ensure unbiasedness under weak conditions on the design.

- **Difference-in-Means**: The difference-in-means estimator is the difference between the averages of the outcomes in the two treatment groups, More formally,

$$\widehat{\tau} = \frac{1}{|Z^+|}\sum_{i\in Z^+} a_i - \frac{1}{|Z^-|}\sum_{i\in Z^-} b_i \ .$$

We remark that under the group-balanced design, the difference-in-means estimator is equivalent to the Horvitz–Thompson estimator. In the next section, we demonstrate typical analyses of an estimator-design pair.

1.2.2 Existing methods of analysis

In this section, we briefly review several central concepts in the statistical analysis of randomized experiments, such as unbiasedness and consistency of estimator-design pairs. These concepts are the bread and butter of the statistician, but may be less familiar to the computer scientist.

Finite Sample Analysis An analysis is said to hold in *finite samples* if it is true for any number of units n. The first finite sample concept we introduce is *unbiasedness*, which ensures that an estimator for the average treatment effect is correct on average.

Definition 1.1. An estimator-design pair is said to be *unbiased* for the average treatment effect if $\mathbb{E}[\widehat{\tau}] = \tau$, where the expectation is with respect to the random assignment specified by the design.

To illustrate the concept of unbiasedness, we show that the Horvitz–Thompson estimator is unbiased for any design satisfying a mild positivity assumption.

Proposition 1.2. *Suppose that a design satisfies* $\Pr(z_i = 1) \in (0, 1)$ *for all units* $i \in [n]$. *Then, the Horvitz–Thompson estimator is an unbiased estimator of the average treatment effect.*

Proof. For each unit $i \in [n]$, define the function $\pi_i(v) = \Pr(z_i = v)$ for $v \in \{\pm 1\}$. Observe that the Horvitz–Thompson estimator may be written as

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{z_i y_i}{\pi_i(z_i)} \ .$$

Under the positivity condition, the expectation of the inner term is

$$\mathbb{E}\Big[\frac{z_i y_i}{\pi_i(z_i)}\Big] = \Pr(z_i = 1) \cdot \frac{1 \cdot a_i}{\pi_i(1)} + \Pr(z_i = -1) \cdot \frac{(-1) \cdot b_i}{\pi_i(-1)} = a_i - b_i = \tau_i \ .$$

The result follows by linearity of expectation and definition of the ATE. □

Unbiasedness is a useful property of an estimator which ensures that there is no systematic bias in our experiment; however, the property itself is weak as it does not speak to the precision of the estimator. The mean squared error, defined below, is a measure of the precision of the estimator.

Definition 1.3. The mean squared error of an estimator-design pair is $\mathbb{E}[(\widehat{\tau} - \tau)^2]$, where the expectation is with respect to the random assignment specified by the design.

One method of analysis is to provide conditions on the potential outcomes under which the mean squared error of an estimator-design pair is small. Such an analysis should inform the experimenter about how to design their experiments. For example, the mean squared error of the Horvitz–Thompson estimator under an arbitrary design is derived in Section 2.1.2 as Lemma 2.2. Bounds on the tails of the distribution of an estimator's error under a design are sometimes derived in the literature to provide further insight into the precision of an experiment.

It is a commonly held (though rarely explicitly stated) belief in the literature that the precision of an estimator-design pair will depend on the potential outcomes in ways that are so involved that succinctly characterizing precision in finite samples is futile. For this reason, statisticians prefer large sample or asymptotic analyses.

Large Sample Analysis An analysis is said to hold in *large samples* if it is a statement about the limiting behavior of an experiment as the number of units grows large. More precisely, these are statements about a sequence of experiments, where the sample size grows to infinity.

Formally speaking, the asymptotic sequence is indexed by $N \in \mathbb{N}$. For every N, the number of units in the N experiment in the sequence is $n = N$, and the outcomes and covariates are doubly indexed $a_i^{(N)}$, $b_i^{(N)}$, and $\boldsymbol{x}_i^{(N)}$, where $i \in [n]$ is the unit. The causal estimand is denoted τ_N and the estimator under the design is denoted $\widehat{\tau}_N$. The experiments in the sequence are not related to each other in any way, but conditions on the sequence are imposed to establish asymptotic properties.

Definition 1.4. An estimator-design pair is *consistent in mean square* for a fixed asymptotic sequence if $\lim_{N\to\infty} \mathbb{E}[(\tau_N - \widehat{\tau}_N)^2] = 0$. Similarly, an estimator-design pair is *N^q-consistent in mean square* if the normalized mean squared error is asymptotically bounded, i.e. $\lim_{N\to\infty} N^q \cdot \sqrt{\mathbb{E}[(\tau_N - \widehat{\tau}_N)^2]} = \mathcal{O}(1)$ for some $q > 0$.

Consistency guarantees that with enough units in the experiment, the estimator-design pair is able to exactly recover the average treatment effect. The parameter q in the Definition 1.4 captures the rate of convergence of the mean squared error to zero. As an example, the next proposition shows that the Horvitz–Thompson estimator is consistent under the Bernoulli design.

Proposition 1.5. *Suppose that the sequence of potential outcomes satisfies the following condition: there exists a constant $C \geq 0$ such that*

$$\frac{1}{n}\sum_{i=1}^{n}(a_i^{(N)} + b_i^{(N)})^2 \leq C \quad \textit{for all} \quad N \in \mathbb{N}.$$

Then, the Horvitz–Thompson estimator under the Bernoulli design is $N^{1/2}$-consistent in mean square.

Proof. We drop the double indexing on the index N of the experiment and consider a fixed number of units n. By Proposition 1.2, the Horvitz–Thompson estimator is

unbiased and so the mean squared error is equal to its variance. Using the form in of the Horvitz–Thompson estimator in the proof of Proposition 1.2, we decompose the variance of the estimator as

$$\mathbb{E}[(\tau - \widehat{\tau})^2] = \mathrm{Var}(\widehat{\tau}) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} 2z_i y_i\Big) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(2z_i y_i, 2z_j y_j) \ .$$

Under the Bernoulli design, the assignments z_i and z_j are independent for $i \neq j$. Because the observed outcomes y_i and y_j are functions of the assignments z_i and z_j, these outcomes are also independent. Thus, the crossing terms in the sum above are zero, yielding that

$$\mathbb{E}[(\tau - \widehat{\tau})^2] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(2z_i y_i) = \frac{1}{n^2}\sum_{i=1}^{n} (a_i - b_i)^2 \ ,$$

where the last equality follows from a computation of these variance terms under the Bernoulli design, i.e.

$$\mathrm{Var}(2z_i y_i) = (1/2)\cdot(2a_i - (a_i - b_i))^2 + (1/2)\cdot(-2b_i - (a_i - b_i))^2 = (a_i + b_i)^2 \ .$$

Under the assumption of the proposition that the average magnitudes of the outcomes are bounded, we have the bound on the mean squared error:

$$\mathbb{E}[(\tau - \widehat{\tau})^2] = \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} (a_i + b_i)^2\right) \leq C/n \ .$$

Thus, we have that the sequence $N^{1/2}\sqrt{\mathbb{E}[(\tau - \widehat{\tau})^2]}$ is asymptotically bounded, as it is upper bounded by the above:

$$\lim_{N\to\infty} N^{1/2} \cdot \sqrt{\mathbb{E}[(\tau - \widehat{\tau})^2]} \leq \lim_{N\to\infty} N^{1/2} \cdot \sqrt{C/N} = C \ . \qquad \square$$

Other asymptotic analyses include characterizing the limiting distribution of an estimator-design pair (e.g. asymptotic normality), which typically require more assumptions on the sequence of potential outcomes than what is required for consistency. For notational convenience, it is common to drop the index N in asymptotic analyses.

An informative way of obtaining asymptotic analyses is to first produce a finite sample bound and then deriving asymptotic conditions under which the finite sample bound yields the desired asymptotic property. Oftentimes, asymptotic analyses can illustrate the key conditions which are required for an estimator to be precise under a design without getting into the details required by a finite sample analysis. Asymptotic analyses of estimator-design pairs are preferred by statisticians in the same way that asymptotic analyses of run time are preferred by computer scientists.

However, asymptotic analyses can often smooth over challenges and difficulties that arise in finite samples. Focusing on the asymptotics can sometimes blind the

methodologist to these nuances—after all, every experiment in the real world includes only finitely many experimental units. Thus, we make an effort throughout this dissertation to carry out finite sample analyses where possible and resort to using asymptotic analyses only for illustrative purposes.

1.2.3 Further remarks on the framework

Before continuing, I would like to highlight and clarify several aspects of the potential outcome framework. My goal in doing so is to better contextualize the framework and its assumptions in the broader fields of causal inference and statistics.

- **Finite Population versus Super-population**: In our description of the framework, there are n units with fixed potential outcomes and covariates. This is referred to as the *finite population* setting. In the *super-population* setting, n units are sampled i.i.d. from a distribution $\mathcal{P}$ over outcome and covariate triples $(a_i, b_i, \boldsymbol{x}_i)$. This places some structural assumptions on the relationship between units, outcomes, and covariates of units in the experiment. The causal estimand is now defined with respect to the distribution (e.g. $\tau = \mathbb{E}_{\mathcal{P}}[a_i - b_i]$) and the randomness in the experiment comes from the experimental design in addition to the underlying distribution on units. The finite population setting is more conservative and sometimes preferred by empirical researchers, as the super-population assumption cannot be verified and may impose a level of homogeneity among units which rarely can be justified.

- **Internal versus External Validity**: The average treatment effect is defined only with respect to the units in the experiment. The ability to estimate the treatment effect to high precision is known as *internal validity*. Even if the experimenter succeeds in the goal of internal validity, this does not address the question of the effect of the treatment on units which were not included in the experiment. Understanding the effect of the treatment on units outside the experiment is known as *external validity*, which is considered by many researchers to be a harder problem. The difference between internal and external validity remains even when a super-population model is assumed (e.g. distribution shifts). Solving the external validity problem is an active area of research (Stuart et al., 2011; Tipton, 2013; Lesko et al., 2017).

- **Other Statistical Problems**: In this dissertation, our primary focus is on point estimation of the average treatment effect. We will also discuss methods for constructing confidence intervals which contain the average treatment effect with a desired probability. There are other statistical problems that one could consider, including various forms of hypothesis testing. Methodological contributions in this work will prove useful for many of these statistical problems, but we focus our attention on point estimation and construction of confidence intervals.

1.2.4 Notation

We denote vectors using lowercase bold text (e.g. $\boldsymbol{x}$ and $\boldsymbol{y}$) and we denote matrices using uppercase bold text (e.g. $\boldsymbol{A}$ and $\boldsymbol{B}$). Coordinates of vectors are denoted using parenthesis so that $\boldsymbol{x}(i)$ denotes the ith component of vector $\boldsymbol{x}$. The coordinates of a matrix are denoted either using the parenthesis notation or a lower-case notation so that $\boldsymbol{A}(i,j)$ or $a_{i,j}$ may be used to denote the entry in the ith row and jth column of $\boldsymbol{A}$. For notational convenience, we do not make distinctions between random variables and the values that they take, which we hope to be clear from context. Given a random vector $\boldsymbol{x}$, we denote its expectation by $\mathbb{E}[\boldsymbol{x}]$ and its covariance matrix by $\mathrm{Cov}(\boldsymbol{x})$.

The transpose of a vector or matrix is denoted $\boldsymbol{x}^\intercal$ and $\boldsymbol{A}^\intercal$, respectively. The Euclidean inner product between vectors is denoted as either $\boldsymbol{x}^\intercal\boldsymbol{y}$ or $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$. The ℓ_2 norm of a vector $\boldsymbol{v}$ is written as $\|\boldsymbol{v}\|$, and the corresponding operator norm of a matrix $\boldsymbol{A}$ is denoted $\|\boldsymbol{A}\| = \max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{A}\boldsymbol{x}\|$. Other norms are introduced throughout the dissertation as they are used.

The trace of a square matrix, denoted $\mathrm{tr}(\boldsymbol{A})$, is the sum of its diagonal entries. We denote the trace inner product on matrices by $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{tr}(\boldsymbol{A}\boldsymbol{B}^\intercal)$. A symmetric matrix $\boldsymbol{A}$ is *positive semidefinite* if $\boldsymbol{v}^\intercal\boldsymbol{A}\boldsymbol{v} \geq 0$ for all vectors $\boldsymbol{v}$ and *positive definite* if this inequality is strict for all vectors. The Loewener ordering is a partial ordering on symmetric matrices defined as $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite.

Chapter 2

Gram–Schmidt Walk Design [1]

2.1 Introduction

Randomized experiments are the gold standard for establishing causal effects and are used in a wide variety of scientific fields. Under the potential outcome framework, randomization of treatment assignments is the cornerstone of estimation and inference in experiments. Exactly *how* to randomize, however, has been the subject of a historical debate which continues to the present day.

Before presenting formal arguments in this debate, consider the following scenario: a researcher aims to estimate the effect of a newly developed drug on a person's heart rate. After recruiting around 100 participants for her study, the researcher sets out to design the randomization scheme used for assigning treatment (either drug or placebo) to the participants. As she investigates randomization schemes, the researcher notices that when treatment is independently assigned to every participant, the two treatment groups tend to look dissimilar in some way: more men are assigned to treatment, the older participants receive control, etc. After careful consideration, the researcher is able to partition the participants into two deterministic groups which are similar in all the ways she can measure. She considers assigning treatment to one of the groups randomly, but worries that there's not enough randomization in this treatment assignment mechanism. The researcher thinks to herself: *should I assign treatment fully at random, or make the treatment groups look similar?*

As stylized as this problem may seem, it is faced by researchers across a variety of disciplines in the design and analysis of their experiments. In Section 2.1.1, we review the debate around randomization and covariate balance in randomized experiments, which dates back to at least 1923 and continues to the present day. In this dissertation chapter, we contribute to this historical debate in two ways.

- Our first contribution is a new perspective on the role of covariate balancing in randomized experiments, which we refer to as the *balance-robustness trade-off*.

[1] Based on the working paper: Christopher Harshaw, Fredrik Sävje, Dan Spielman, and Peng Zhang (2021) "Balancing covariates in randomized experiments with the Gram–Schmidt Walk design". arXiv:1911.03071.

As we discuss in more detail later in the Chapter, we say that a design is robust if the estimator is sufficiently precise for all possible values of the unknown potential outcomes, which is a worst-case analysis. We highlight and formalize the following fundamental tension: in finite samples, an experiment cannot balance covariates and provide maximal robustness. Understanding the effect of this trade-off on the results of an experiment is a statistical question, while constructing an experimental design which achieves a desired trade-off between covariate balance and robustness is an inherently computational question.

- Our second contribution is the development of the Gram–Schmidt Walk Design (GSW-Design), which allows experimenters to navigate the balance-robustness trade-off in a near-optimal way. We obtain several tight characterizations of the behavior of the Horvitz–Thompson estimator under the GSW-Design—including bounds on the mean squared error and tails of the estimator—which hold in finite samples and without any structural assumptions on the units, outcomes, or covariates. Our finite-sample analysis gives a clearer understanding about the way in which covariate balancing may improve precision of the Horvitz–Thompson estimator.

Central to the design and its analysis is the field of *algorithmic discrepancy theory*, which seeks to answer the following questions: how well can a group of objects be partitioned into two similar groups? Are there efficient algorithms for constructing such partitions? The objects considered by discrepancy theory range from geometric (e.g. points in a square or discs in the plane) to combinatorial (e.g. set systems) and even linear algebraic (e.g. vectors and linear operators) in nature (Beck and Fiala, 1981; Spencer, 1985; Banaszczyk, 1998; Marcus et al., 2015). For each setting, a measure of discrepancy is proposed and algorithms for constructing a partition with small discrepancy are considered. Indeed, our design is based on the Gram–Schmidt Walk algorithm of Bansal et al. (2019), which provided the first efficient algorithm for balancing high dimensional vectors in a way that constructively proves Banaszczyk's theorem.

One contribution of this chapter is the proposal of a new *distributional discrepancy* problem (Section 2.2.3, Problem 2.5) which, unlike previous discrepancy problems that focus only a single partition, is defined for a distribution over partitions. This distributional discrepancy problem is more well-suited to the design of randomized experiments. To the best of our knowledge, this is the first formal connection between the two disparate fields: the design of randomized experiments and algorithmic discrepancy theory.

A deep understanding of discrepancy theory is not required to understand the results presented in this chapter. Unfortunately, it is well beyond the scope of this chapter to provide a complete overview of discrepancy theory. In order to get a sense of the mathematical foundation underpinning this work, we point the reader to the references above, the textbooks (Chazelle, 2000; Matoušek, 1999) and the book chapter of Bansal (2014).

2.1.1 Related work

We begin by reviewing the the historical debate on the use of randomization and covariate balancing in experimental design. In favor of randomization, Fisher (1925, 1926) points out that randomizing ensures unbiased estimates of the treatment effect, but his arguments also extend to precision of an estimator. Wu (1981) appears to be the first paper to explicitly discuss the connection between randomization and robustness in this extended sense. He shows that the fully randomized design minimizes the worst-case mean squared error. The result has been extended in various directions (e.g., Li, 1983; Kallus, 2018; Bai, 2019b; Basse et al., 2019).

In a review of the experimental methods of the day, Student (1923) did not mention randomization even as a possibility. In a later paper, Student (1938) explicitly argues that randomization often is harmful because random assignments can only make the treatment groups less comparable than what they would be under the most balanced assignment. His conclusion is that the only role for randomization is to select between assignments that are equally balanced. The same conclusion, in slightly different incarnations, has been reached several times after this (see, e.g., Bertsimas et al., 2015; Kasy, 2016; Deaton and Cartwright, 2018; Kallus, 2018).

Fisher highlights that we do not need to choose between the two extremes. We can partially restrict the randomization to avoid the most troublesome imbalances, but allow some imbalances to persist in order to accommodate well-motivated confidence intervals and hypothesis tests. The insight has inspired a large number of experimental designs which fall on the continuum between the fully randomized and the maximally balanced designs. Examples include the matched pair design (Greevy et al., 2004; Imai et al., 2009; Bruhn and McKenzie, 2009), various stratified designs (Fisher, 1935; Higgins et al., 2016) and re-randomization (Lock Morgan and Rubin, 2012; Li et al., 2018).

We now discuss two types of covariate-balancing experimental designs which are most relevant to the results presented in this chapter: re-randomization and matched pair design.

A *re-randomization design* is a uniform distribution over the set of assignments which satisfy a user-defined balance criterion. Experimenters use rejection sampling to sample from a re-randomization design; that is, assignment vectors are independently and uniformly sampled until one meets the balance criterion. Lock Morgan and Rubin (2012) provide an analysis of this approach when the balance criterion is based on Mahalanobis distances: an assignment $\boldsymbol{z} \in \{\pm 1\}^n$ is accepted if

$$\left\| \sum_{i \in Z^+} \boldsymbol{v}_i - \sum_{i \in Z^-} \boldsymbol{v}_i \right\| \leq \alpha \quad \text{where for each unit } i \in [n],\ \boldsymbol{v}_i = \left(\sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^\intercal \right)^{-1/2} \boldsymbol{x}_i$$

is a whitened-covariate vector and $\alpha > 0$ is a user-defined threshold. They investigate the improvement in precision achieved by the approach under an assumption of additive treatment effects and an assumption of normally distributed covariates

and potential outcomes. Under these assumptions, they show that rerandomization monotonically improves precision in the estimator as the balance criterion becomes stricter. However, the assumptions they impose implicitly remove the trade-off between robustness and balance mentioned above, and it may therefore be difficult for experimenters to judge the relevance of the results for practice. Rerandomization also suffers from computational difficulties: to decrease the balance parameter α by a constant factor one must reject a number of samples that is exponential in the number of covariates.

Li et al. (2018) relax the assumptions imposed by Lock Morgan and Rubin (2012) at the cost of studying the estimator's asymptotic distribution. The authors show that the estimator has a non-normal asymptotic distribution under rerandomization in this more general setting, emphasizing the concerns with the assumptions in the analysis by Lock Morgan and Rubin (2012). While Li et al. improve understanding of the rerandomization design, their analysis considers a balance criterion that is asymptotically fixed, meaning that an acceptable assignment must attain an aggregated Mahalanobis distance below some fixed threshold no matter the sample size. In practice, experimenters are typically limited to those acceptance criterion which their computational resources allow, which typically varies by sample size and dimensionality of the covariates. It may therefore be difficult for experimenters to use the asymptotic results in practice.

Another related design is the *matched pair design*, which consists of two parts. First, the units are grouped into pairs based upon the similarity of their covariates. Next, independently for each pair, one unit is randomly selected to receive treatment and the other unit receives control. The matched pair design itself is agnostic to the way in which units are matched, although Greevy et al. (2004) advocate for using a min-cost matching formulation where the cost of a matched pair is the Euclidean distance between (possibly whitened) covariate vectors. One downside of the matching-based approach is that high dimensional covariates can be nearly equidistant and so most matchings based on pair-wise distances may be indistinguishable.

Recently, Bai (2019a) analyzed the matched pair design under a super-population framework, where the outcomes and covariates of each unit are drawn i.i.d. from an unknown distribution. Bai (2019a) shows that among all stratified designs, a matched pair design minimizes the mean squared error. This optimality analysis has two main drawbacks. First, the matched pair design is shown to minimize mean squared error only among stratified designs; however, this optimality will not hold among general designs. Second, the optimal matching depends on the unknown outcomes, which are not available to the experimenter. Bai (2019a) provides methods for estimating the optimal matching when results from a pilot study are available. It is possible that similar pilot study considerations may be used to inform the choice of parameter for the GSW-Design presented here.

Compared to analyses of these previous designs, the analysis in this chapter does not require assumptions on the potential outcomes or the covariates. In particular,

the analysis does not require the covariates or potential outcomes to be normally distributed, nor the units to be drawn from a superpopulation, nor the treatment effects to be additive. Moreover, the analysis does not require whitening transformations on the covariates, so it applies no matter what type of covariates experimenters want to balance. Still, the analysis of both precision and tail behavior is valid in finite samples, and does not rely on large sample approximations. We use asymptotic illustrations only for expositional purposes to highlight and simplify features of the finite sample results. Hence, the understanding of the behavior of the Gram–Schmidt Walk design is both more precise and more relevant to practice than the understanding of existing experimental designs. Additionally, whereas some existing designs require excessive computational resources, sampling assignments from the Gram–Schmidt Walk design is practical due to its computationally efficiency.

2.1.2 Preliminaries

We work under the Neyman-Rubin model of potential outcomes, which we briefly summarize here. There are n units in the experiment, indexed by $[n] = \{1, 2, \dots n\}$. The experimenter randomly assigns treatment $z_i \in \{\pm 1\}$ to each unit $i \in [n]$ and we collect these assignments into the random vector $\boldsymbol{z} = (z_1, z_2, \dots z_n)$. We denote

$$Z^+ = \{i \in [n] : z_i = 1\} \quad \text{and} \quad Z^- = \{i \in [n] : z_i = -1\}$$

to be the random partition of the units into treatment and control groups, respectively. The *design* of the experiment is the distribution over the assignment vectors $\{\pm 1\}^n$.

Each unit $i \in [n]$ has two potential outcomes, a_i which is observed if $z_i = 1$ and b_i which is observed if $z_i = -1$. The term potential is used here because while both outcomes have the potential to be observed, only one of them is. For each unit $i \in [n]$, the observed outcome is the random variable defined by

$$y_i = \begin{cases} a_i & \text{if } z_i = 1 \\ b_i & \text{if } z_i = -1 \end{cases}$$

It will be convenient to collect the outcome variables into vectors

$$\boldsymbol{a} = (a_1, a_2, \dots a_n) \quad \boldsymbol{b} = (b_1, b_2, \dots b_n) \quad \boldsymbol{y} = (y_1, y_2, \dots y_n) \ .$$

Each unit $i \in [n]$ has a vector of d covariates, $\boldsymbol{x}_i \in \mathbb{R}^d$ which is known to the experimenter prior to treatment assignment. In this way, the design may depend on the pre-treatment covariates. We emphasize that units, their outcomes, and their covariates are deterministic and we make no assumptions on them; the only randomness is the experiment comes from the experimenter's assignment of treatment.

The causal quantity of interest is the *average treatment effect*

$$\tau = \frac{1}{n}\sum_{i=1}^{n}(a_i - b_i) \ ,$$

which is the contrast between outcomes under treatment and control, averaged over units in the experiment. The average treatment effect cannot be directly observed and so we must estimate it. In this chapter, we restrict our attention to the Horvitz–Thompson estimator

$$\widehat{\tau} = \frac{1}{n}\left[\sum_{i \in Z^+} \frac{a_i}{\Pr(z_i = 1)} - \sum_{i \in Z^-} \frac{b_i}{\Pr(z_i = -1)}\right] \ .$$

Proposition 1.2 in Section 1.2 demonstrates that the Horvitz–Thompson estimator is unbiased under designs which satisfy the positivity condition that $\Pr(z_i = 1) \in (0, 1)$ for all units $i \in [n]$.

When comparing designs, we focus on the precision of the Horvitz–Thompson estimator. To make the task concrete, we investigate the mean squared error $\mathbb{E}[(\tau - \widehat{\tau})^2]$. For expositional purposes, we restrict our attention throughout the chapter to designs where each unit is equally likely to receive either treatment, i.e. $\Pr(z_i = 1) = \Pr(z_i = -1) = 1/2$. Extensions of our results to settings where $\Pr(z_i = 1) \in (0, 1)$ are discussed in Appendix A.1.6.

The following lemma derives the error of the Horvitz–Thompson estimator conditioned on a particular assignment.

Lemma 2.1. *For any experimental design satisfying* $\Pr(z_i = 1) = 1/2$ *for all units* $i \in [n]$*, the error of the Horvitz–Thompson estimator can be written as*

$$\widehat{\tau} - \tau = \frac{1}{n}\langle \boldsymbol{z}, \boldsymbol{\mu} \rangle \qquad \textit{where} \qquad \boldsymbol{\mu} = \boldsymbol{a} + \boldsymbol{b}.$$

Proof. Recall that the average treatment effect and Horvitz–Thompson estimator can be written as

$$\tau = \frac{1}{n}\langle \mathbf{1}, \boldsymbol{a} - \boldsymbol{b} \rangle \qquad \text{and} \qquad \widehat{\tau} = \frac{2}{n}\langle \boldsymbol{z}, \boldsymbol{y} \rangle.$$

By expressing the observed outcome as $y_i = a_i(\frac{1+z_i}{2}) + b_i(\frac{1-z_i}{2})$, we see that

$$n\widehat{\tau} = 2\langle \boldsymbol{z}, \boldsymbol{y} \rangle = \langle \boldsymbol{z}, \boldsymbol{a} + \boldsymbol{b} \rangle + \langle \mathbf{1}, \boldsymbol{a} - \boldsymbol{b} \rangle = \langle \boldsymbol{z}, \boldsymbol{\mu} \rangle + n\tau.$$

The desired result is obtained by rearranging terms. □

Lemma 2.1 demonstrates that the mean squared error depends on the potential outcomes through their sum. For this reason, we refer to $\boldsymbol{\mu}$ as the *sum potential outcome vector*, or simply the *potential outcome vector* in this chapter. We now derive the mean squared error of the Horvitz–Thompson estimator under an arbitrary

design.

Lemma 2.2. *For any experimental design with* $\Pr(z_i = 1) = 1/2$ *for all* $i \in [n]$*, the mean squared error of the Horvitz–Thompson estimator is*

$$\mathbb{E}[(\widehat{\tau} - \tau)^2] = \frac{1}{n^2}\boldsymbol{\mu}^\intercal \operatorname{Cov}(\boldsymbol{z})\boldsymbol{\mu} \ .$$

Proof. Lemma 2.1 gives $\widehat{\tau} - \tau = \langle \boldsymbol{z}, \boldsymbol{\mu} \rangle / n$. The expectation of the square of this expression is

$$\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] = \frac{1}{n^2}\,\mathbb{E}\big[\langle \boldsymbol{z}, \boldsymbol{\mu} \rangle^2\big] = \frac{1}{n^2}\boldsymbol{\mu}^\intercal\,\mathbb{E}\big[\boldsymbol{z}\boldsymbol{z}^\intercal\big]\boldsymbol{\mu},$$

because $\boldsymbol{\mu}$ is not random. The proof is completed by noting that $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal] = \operatorname{Cov}(\boldsymbol{z})$ because $\mathbb{E}[\boldsymbol{z}] = \mathbf{0}$ when $\Pr(z_i = 1) = 1/2$ for all $i \in [n]$. □

Lemma 2.2 demonstrates that the mean squared error of the Horvitz–Thompson estimator under a given design is the quadratic form in $\operatorname{Cov}(\boldsymbol{z})$ evaluated at the (unknown) potential outcome vector $\boldsymbol{\mu}$. In this way, the properties of the design which affect the mean squared error are completely captured by the covariance matrix of assignments, $\operatorname{Cov}(\boldsymbol{z})$. This is a central insight in our work which informs both our interpretation of the experimental design problem as well as the proposed design.

2.2 A New Perspective on Covariate Balancing

In this section, we present new notions of robustness and covariate balance in randomized experiments. Our notion of covariate balance is motivated by a presumed linear relation between outcomes and covariates. We show that there is a fundamental trade-off between covariate balance and robustness.

2.2.1 A measure of robustness

When designing an experiment, the experimenter is often primarily concerned with the *robustness* of the experiment. Informally speaking, we say that a design is robust if the estimator is sufficiently precise for all possible values of the unknown potential outcomes. This can be formalized by examining the maximum mean squared error over a particular set of potential outcomes. In this way, the design is robust if this worst-case error is not too large. An experimenter may value a design which is robust, as the potential outcomes are unknown before experimentation.

Our first result is to show that the operator norm of the covariance matrix of treatment assignments $\operatorname{Cov}(\boldsymbol{z})$ characterizes the worst-case behavior of the design over all potential outcomes of bounded average magnitude. Here, the set of potential outcomes with bounded average magnitude is defined as

$$\mathrm{PO}(M) = \left\{ \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^{n} (a_i + b_i)^2 \leq M \right\} \ ,$$

where M is the bound on the average magnitude.

Lemma 2.3. *Consider a design satisfying* $\Pr(z_i = 1) = 1/2$ *for all units* $i \in [n]$*. The worst-case mean squared error over the set of all potential outcomes with bounded magnitude is*

$$\max_{\boldsymbol{a},\boldsymbol{b} \in PO(M)} \mathbb{E}[(\tau - \widehat{\tau})^2] = \frac{M}{n} \cdot \|\mathrm{Cov}(\boldsymbol{z})\|.$$

Lemma 2.3 motivates taking the operator norm[2] $\|\mathrm{Cov}(\boldsymbol{z})\|$ to be a measure of robustness in a randomized experiment, as it determines the worst-case error of the Horvitz–Thompson estimator. The operator norm $\|\mathrm{Cov}(\boldsymbol{z})\|$ measures the amount of pair-wise dependence between the assignments. Thus, Lemma 2.3 demonstrates that designs with greater amounts of dependence between assignments are less robust.

Our next result is to show that the Bernoulli design is min-max optimal over the set of potential outcomes with bounded average magnitude. The result follows by observing that the Bernoulli design minimizes the operator norm $\|\mathrm{Cov}(\boldsymbol{z})\|$ over all designs satisfying $\Pr(z_i = 1) = 1/2$.

Proposition 2.4. *Every experimental design with* $\Pr(z_i = 1) = 1/2$ *for all units* $i \in [n]$ *satisfies the inequality* $\|\mathrm{Cov}(\boldsymbol{z})\| \geq 1$ *and equality holds for the Bernoulli design. Thus, the Bernoulli design is min-max optimal for potential outcomes with bounded average magnitude,* $PO(M)$*.*

Proposition 2.4 suggests that the experimenter who seeks maximal amounts of robustness should employ the Bernoulli design, where treatment assignments are maximally independent. Another implication of these results is that the operator norm $\|\mathrm{Cov}(\boldsymbol{z})\|$ measures the multiplicative increase in the mean squared error from that of the min-max design.

2.2.2 A measure of covariate balance

Oftentimes, the experimenter has prior substantive knowledge about the units and their outcomes. In particular, the pre-treatment covariates might inform the experimenter about which units will have similar outcomes. In this case, the experimenter may wish to forgo some amount of robustness in order to increase precision for certain outcomes. When covariates are somewhat informative of the outcomes, a design which ensures *covariate balance* between the treatment groups may help to increase precision.

We formalize this covariate balancing intuition by considering the best linear association between outcomes and covariates. Recall that each unit $i \in [n]$ has an associated d-dimensional vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^d$. We collect the covariate vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n$ as rows of the n-by-d matrix $\boldsymbol{X}$. Let $\boldsymbol{\beta}_{\text{LS}}$ be a best linear fit of the

[2]Recall from Section 1.2.4 that the operator norm is defined as $\|\mathrm{Cov}(\boldsymbol{z})\| = \max_{\|\boldsymbol{w}\|=1} \|\mathrm{Cov}(\boldsymbol{z})\boldsymbol{w}\|$.

outcomes to the covariates:

$$\boldsymbol{\beta}_{\text{LS}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\| \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta} \right\| \ .$$

We emphasize that $\boldsymbol{\beta}_{\text{LS}}$ is not known to the experimenter and has no causal interpretation—it is simply the projection of the outcome vector onto the span of the covariates. We may decompose the outcome vector into two orthogonal parts,

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}} + (\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}) \triangleq \widehat{\boldsymbol{\mu}} + \boldsymbol{\varepsilon} \ ,$$

where $\widehat{\boldsymbol{\mu}}$ is the best linear fit of the outcomes to the covariates and $\boldsymbol{\varepsilon}$ is the residual error of this fit. Using Lemma 2.2, we may write the mean squared error of a design in terms of this decomposition as

$$\begin{aligned}
\mathbb{E}[(\tau - \widehat{\tau})^2] &= \boldsymbol{\mu}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{z}) \boldsymbol{\mu} \\
&= (\widehat{\boldsymbol{\mu}} + \boldsymbol{\varepsilon})^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{z}) (\widehat{\boldsymbol{\mu}} + \boldsymbol{\varepsilon}) \\
&= \widehat{\boldsymbol{\mu}}^{\mathsf{T}} \operatorname{Cov}(z) \widehat{\boldsymbol{\mu}} + 2\widehat{\boldsymbol{\mu}} \operatorname{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta}_{\text{LS}}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{z}) \boldsymbol{\beta}_{\text{LS}} + 2\boldsymbol{\beta}_{\text{LS}} \boldsymbol{X}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon}
\end{aligned}$$

Ignoring the crossing term for the moment, we have that the mean squared error of a design is roughly

$$\mathbb{E}[(\tau - \widehat{\tau})^2] \approx \boldsymbol{\beta}_{\text{LS}}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{z}) \boldsymbol{\beta}_{\text{LS}} + \boldsymbol{\varepsilon}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} \ .$$

Suppose that the covariates are highly linearly predictive of the potential outcomes so that $\widehat{\boldsymbol{\mu}}$ has a considerably larger norm than the residual $\boldsymbol{\varepsilon}$. In this case, we can make the mean squared error small by aiming to make the $\boldsymbol{\beta}_{\text{LS}}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{z}) \boldsymbol{\beta}_{\text{LS}}$ term small. The best-fit linear function $\boldsymbol{\beta}_{\text{LS}}$ is not known, so one way to ensure that this term is small is to use the operator-norm bound:

$$\boldsymbol{\beta}_{\text{LS}}^{\mathsf{T}} \operatorname{Cov}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{z}) \boldsymbol{\beta}_{\text{LS}} \leq \|\operatorname{Cov}(\boldsymbol{X}\boldsymbol{z})\| \cdot \|\boldsymbol{\beta}_{\text{LS}}\|^2$$

and construct the design so that the operator norm $\|\operatorname{Cov}(\boldsymbol{X}\boldsymbol{z})\|$ is small. This ensures that the mean squared error is small for all potential outcome vectors which are well-approximated by a linear function of the covariates, regardless of *which* linear function it is. For this reason, we consider $\|\operatorname{Cov}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{z})\|$ to be an informative notion of covariate balance for the design.

An experimenter may now wish to construct a design which is robust and achieves a high level of covariate balance. The fundamental trade-off is that a design cannot maximally balance covariates and attain maximal robustness: both operator norms $\|\operatorname{Cov}(\boldsymbol{z})\|$ and $\|\operatorname{Cov}(\boldsymbol{X}^{\mathsf{T}} \boldsymbol{z})\|$ cannot be simultaneously minimized. A maximally robust design requires nearly uncorrelated treatment assignments while a maximally covariate-balancing design requires a large amount of correlation between treatment

assignments. Generally speaking, a design which achieves increased covariate balance will be less robust and an extremely robust design affords little room for balancing covariates. We refer to this fundamental tension as the *balance robustness trade-off*.

How to navigate the balance-robustness trade-off depends on the preferences of the experimenter. Some experimenters may prefer increased covariate balance at the cost of some robustness, while other experimenters may value robustness over any level of covariate balance. The goal in the remainder of the chapter is to provide algorithmic insights and techniques into how this balance-robustness trade-off may be navigated.

2.2.3 A distributional discrepancy problem

In this section, we introduce a new algorithmic discrepancy problem, which captures the balance-robustness trade-off and arises naturally in the context of experimental design. Unlike most previously considered discrepancy problems which seek to construct a single assignment vector, this is a *distributional discrepancy* problem, which seeks to construct a distribution over assignment vectors.

Problem 2.5. *Consider input vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_n \in \mathbb{R}^d$ arranged as rows in the n-by-d matrix $\boldsymbol{X}$. What is the smallest value $C > 0$ such that there exists an efficient randomized algorithm which takes as input $\boldsymbol{X}$ and $\phi \in [0, 1]$ and produces an assignment $\boldsymbol{z} \in \{\pm 1\}^n$ with the following distributional properties:* $\Pr(z_i = 1) = 1/2$ *for all* $i \in [n]$,

$$\|\mathrm{Cov}(\boldsymbol{z})\| \leq \frac{1}{\phi} \quad \textit{and} \quad \frac{1}{\xi^2}\|\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})\| \leq \frac{C}{1-\phi} \ ,$$

where $\xi = \max_{i \in [n]} \|\boldsymbol{x}_i\|$ is the maximum norm of the input vectors?

As discussed in the two sections above, the first norm captures the robustness and the second norm captures the covariate balance. Problem 2.5 captures the trade-off between these quantities through the parameter $\phi \in [0, 1]$. As $\phi \to 1$, all emphasis is placed on robustness and when $\phi \to 0$, all emphasis is placed on covariate balance. The maximum norm $\xi = \max_{i \in [n]} \|\boldsymbol{x}_i\|$ appears as a scaling factor so that the problem remains unaffected by scaling of the covariates, i.e. $\alpha \cdot \boldsymbol{X}$ for a scalar $\alpha \in \mathbb{R}$.

We now highlight the way in which Problem 2.5 is a discrepancy problem, albeit a distributional one. Given a set of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_n \in \mathbb{R}^d$, the *discrepancy vector* of an assignment $\boldsymbol{z} \in \{\pm 1\}^n$ is the difference of within-group sums:

$$\boldsymbol{X}^\intercal \boldsymbol{z} = \sum_{i=1}^{n} z_i \boldsymbol{x}_i = \sum_{i \in Z^+} \boldsymbol{x}_i - \sum_{i \in Z^-} \boldsymbol{x}_i \ .$$

The *discrepancy* of an assignment is a measurement of the magnitude of the corresponding discrepancy vector, typically with the squared Euclidean norm or the infinity norm. The squared Euclidean norm may be expressed in the following variational

way:

$$\|\boldsymbol{X}^\intercal \boldsymbol{z}\|^2 = \max_{\substack{\boldsymbol{\theta}\in\mathbb{R}^d \\ \|\boldsymbol{\theta}\|=1}} \langle \boldsymbol{\theta}, \boldsymbol{X}^\intercal \boldsymbol{z}\rangle^2 = \max_{\substack{\boldsymbol{\theta}\in\mathbb{R}^d \\ \|\boldsymbol{\theta}\|=1}} \Big(\sum_{i\in Z^+} \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle - \sum_{i\in Z^-} \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle \Big)^2$$

We now show that Problem 2.5 has the same interpretation, but we are interested in the *distribution* of the random discrepancy vector. Using the definition of the operator norm and expanding terms, we can write the operator norm $\|\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})\|$ in a similarly variational form,

$$\|\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})\| = \max_{\substack{\boldsymbol{\theta}\in\mathbb{R}^d \\ \|\boldsymbol{\theta}\|=1}} \mathbb{E}\Big[\Big(\sum_{i\in Z^+} \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle - \sum_{i\in Z^-} \langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle \Big)^2 \Big] \ .$$

The equality above demonstrates that the operator norm $\|\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})\|$ is the maximum expected squared imbalance between the two partitions, as measured by a linear function of the discrepancy vector. In this sense, $\|\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})\|$ may be understood as the distributional extension of the squared Euclidean discrepancy. Of course, the key aspect of Problem 2.5 is the trade-off between this and the correlation of assignments $\|\mathrm{Cov}(\boldsymbol{z})\|$ which we have expressed above in a variational way.

We emphasize here that Problem 2.5 is not subsumed or solved by previously considered discrepancy problems in the literature. Indeed, the goal of most discrepancy problems is to produce a *single* assignment vector $\boldsymbol{z}$ which minimizes a norm of the discrepancy vector. A naive application of discrepancy minimization to Problem 2.5 is to choose $\boldsymbol{z}^*$ to be the assignment which minimizes the squared Euclidean norm $\|\boldsymbol{X}^\intercal \boldsymbol{z}\|^2$ and construct a distribution by choosing either $\boldsymbol{z}^*$ or $-\boldsymbol{z}^*$ with equal probability. This naive experimental design may result in substantial covariate balance, as $\|\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})\| = \|\boldsymbol{X}^\intercal \boldsymbol{z}\|^2$; however, this design affords virtually no robustness, as it yields $\|\mathrm{Cov}(\boldsymbol{z})\| = n$. Thus, Problem 2.5 is a truly new discrepancy problem which requires new insights and algorithmic considerations.

The only other distributional discrepancy problem that we are aware of is the subgaussian discrepancy problem introduced by Dadush et al. (2019), which led to the development of the Gram–Schmidt Walk algorithm of Bansal et al. (2019). This discrepancy problem is similar to ours when considering only $\phi = 0$.

In Problem 2.5, the relevant value of C is a constant, independent of the number of input vectors n and their dimension d. We remark that C must be at least 1 for Problem 2.5 to be solvable. For example, $\xi^{-2}\|\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})\|$ cannot be made smaller than 1 for orthogonal vectors when $\Pr(z_i = 1) = 1/2$ for all units $i \in [n]$.

2.2.4 Implications for mean squared error analysis

Finally, we demonstrate that an algorithm which solves Problem 2.5 yields an experimental design under which we can reason about the mean squared error of the Horvitz–Thompson estimator.

The first insight is that the parameter $\phi \in [0, 1]$ controls the robustness of the

design. In particular, the requirements of Problem 2.5 specify that $\|\mathrm{Cov}(\boldsymbol{z})\| \leq 1/\phi$. As discussed in Section 2.2.1, this controls the increase in the worst-case mean squared error over the max-min design. In this sense, a maximally robust design is achieved for $\phi = 1$ and less robustness is offered as ϕ decreases.

The second insight is that decreasing ϕ can lead to a provable decrease in mean squared error of the Horvitz–Thompson estimator when the covariates are (at least somewhat) linearly predictive of the outcomes. This is formalized below as Theorem 2.6.

Theorem 2.6. *Suppose that $\mathcal{A}$ is an algorithm which satisfies the requirements of Problem 2.5, taking $\boldsymbol{X}$ and $\phi \in [0, 1]$ as input. Then, the mean squared error of the Horvitz–Thompson estimator under the experimental design given by $\mathcal{A}$ is bounded by the loss of the ridge regression:*

$$\mathbb{E}[(\tau - \widehat{\tau})^2] \leq \frac{1}{n} \cdot \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left[\frac{1}{\phi^2} \frac{1}{n} \sum_{i=1}^{n} \Big((a_i + b_i) - \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle \Big)^2 + \frac{C\xi^2}{(1-\phi)^2 n} \|\boldsymbol{\beta}\|^2 \right] .$$

Proof. Let $\boldsymbol{\beta} \in \mathbb{R}^d$ be an arbitrary vector and let $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon} = \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}$ so that $\boldsymbol{\mu} = \boldsymbol{\varepsilon} + \widehat{\boldsymbol{\mu}}$. First, we use Lemma 2.2 together with a generalized arithmetic-geometric (AM-GM) inequality to separate the mean squared error into two parts: one which depends on the linear prediction $\widehat{\boldsymbol{\mu}}$ and the other which depends on the residual $\boldsymbol{\varepsilon}$. For all $\gamma > 0$,

$$\begin{aligned}
n^2\, \mathbb{E}[(\tau - \widehat{\tau})^2] &= \boldsymbol{\mu}^\mathsf{T} \mathrm{Cov}(\boldsymbol{z}) \boldsymbol{\mu} \\
&= (\widehat{\boldsymbol{\mu}} + \boldsymbol{\varepsilon})^\mathsf{T} \mathrm{Cov}(\boldsymbol{z}) (\widehat{\boldsymbol{\mu}} + \boldsymbol{\varepsilon}) \\
&= \widehat{\boldsymbol{\mu}}^\mathsf{T} \mathrm{Cov}(\boldsymbol{z}) \widehat{\boldsymbol{\mu}} + \boldsymbol{\varepsilon}\, \mathrm{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} + 2\widehat{\boldsymbol{\mu}}^\mathsf{T} \mathrm{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} \\
&\leq (1 + \gamma^2)\ \widehat{\boldsymbol{\mu}}^\mathsf{T} \mathrm{Cov}(\boldsymbol{z}) \widehat{\boldsymbol{\mu}} + (1 + \gamma^{-2})\ \boldsymbol{\varepsilon}^\mathsf{T} \mathrm{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} \\
&= (1 + \gamma^2)\ \boldsymbol{\beta}^\mathsf{T} \mathrm{Cov}(\boldsymbol{X}^\mathsf{T}\boldsymbol{z}) \boldsymbol{\beta} + (1 + \gamma^{-2})\ \boldsymbol{\varepsilon}^\mathsf{T} \mathrm{Cov}(\boldsymbol{z}) \boldsymbol{\varepsilon} \ .
\end{aligned}$$

By applying the operator norm bound to the quadratic forms and using the bounds on the two operator norms guaranteed by Problem 2.5, we have that

$$\begin{aligned}
&\leq (1 + \gamma^2)\ \|\mathrm{Cov}(\boldsymbol{X}^\mathsf{T}\boldsymbol{z})\| \|\boldsymbol{\beta}\|^2 + (1 + \gamma^{-2})\ \|\mathrm{Cov}(\boldsymbol{z})\| \|\boldsymbol{\varepsilon}\|^2 \\
&\leq (1 + \gamma^2) \Big(\frac{\xi^2 C}{1 - \phi} \Big) \|\boldsymbol{\beta}\|^2 + (1 + \gamma^{-2}) \frac{1}{\phi} \|\boldsymbol{\varepsilon}\|^2 \\
&= \frac{C\xi^2}{(1-\phi)^2} \|\boldsymbol{\beta}\|^2 + \frac{1}{\phi^2} \|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \ ,
\end{aligned}$$

where the final equality follows by setting $\gamma^2 = \phi/(1 - \phi)$, and recalling that $\boldsymbol{\varepsilon} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}$. Note that this upper bound holds for an arbitrary vector $\boldsymbol{\beta}$, and so the result follows by minimizing over all such $\boldsymbol{\beta}$. □

Theorem 2.6 demonstrates the relevance of Problem 2.5 to the experimental design

problem and formalizes the balance-robustness trade-off in terms of the precision of the estimator. Theorem 2.6 demonstrates that the mean squared error depends on the parameter ϕ and the degree to which the covariates are predictive of the potential outcomes. The predictiveness is captured by the minimum loss of a ridge regression of the potential outcomes on the covariates. Note that the Horvitz–Thompson estimator itself does not conduct any covariate adjustment, instead being a raw comparison of outcomes between treatment groups. Indeed, the ridge regression is never actually run and cannot be performed by the experimenter, as it depends on all potential outcomes. In this sense, we say that the designs satisfying the requirements of Problem 2.5 conduct *regression by design*.

The first term in the upper bound of Theorem 2.6 captures how well a linear function $\boldsymbol{\beta}$ predicts the potential outcomes using the covariates. This term can be made small if the potential outcome vector is close to the span of the covariates. The second term captures the magnitude of the linear function, as measured by the sum of the squares of the coefficients. The factor ξ^2 puts this magnitude on a neutral scale so that the optimum is not affected by a rescaling of the covariates. The design parameter ϕ determines the trade-off between the two terms, assigning more focus to either finding a function that predicts the outcomes well or one that is of small magnitude. Put differently, the theorem tells us that the design performs well when the potential outcomes can be well approximated by a relative simple linear function of the covariates, as measured by coefficient norm.

In theory, an experimenter might want to resolve the balance-robustness trade-off by choosing the trade-off parameter ϕ^* which minimizes the upper bound on the mean squared error:

$$\phi^* = \left(1 + \sqrt{C}\frac{\xi\|\boldsymbol{\beta}_{\text{LS}}\|}{\|\boldsymbol{\varepsilon}\|}\right)^{-1} .$$

However, this is not possible because ϕ^* depends on the potential outcomes, which are unknown to the experimenter before running the experiment. It is beyond the scope of this dissertation to present a formal framework for how experimenters ought to navigate the balance-robustness trade-off, although we give some recommendations for setting ϕ in Section 2.4.2.

We remark that although Theorem 2.6 holds for any sampling algorithm which satisfies the requirements of Problem 2.5, improved bounds on the mean squared error may be obtained for specific sampling algorithms. In particular, the AM-GM inequality introduces some loss which may be avoided by more closely analyzing the crossing terms. The remainder of the chapter is devoted to describing and analyzing the GSW-Design, which solves the distributional discrepancy problem (Problem 2.5) and for which we can obtain an even tighter analysis than that guaranteed by Theorem 2.6.

2.3 The Gram–Schmidt Walk Design

In this section, we present the Gram–Schmidt Walk Design (GSW-Design) for navigating the balance-robustness trade-off, as formulated in Problem 2.5. At a high level, GSW-Design operates by attempting to balance *augmented covariate vectors*. For each unit $i \in [n]$, we define the augmented covariate vector $\boldsymbol{b}_i \in \mathbb{R}^{n+d}$ to be a scaled concatenation of the unit's raw covariate and a unit-unique indicator variable:

$$\boldsymbol{b}_i = \begin{bmatrix} \sqrt{\phi}\boldsymbol{e}_i \\ \xi^{-1}\sqrt{1-\phi}\boldsymbol{x}_i \end{bmatrix},$$

where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the ith basis vector of dimension n. As $\phi \to 1$, the augmented covariate vectors become a set of orthogonal vectors and as $\phi \to 0$, the augmented covariate vectors begin to more closely resemble the original raw covariate vectors.

The GSW-Design constructs the augmented covariate vectors and uses them as input to the Gram–Schmidt Walk Algorithm of Bansal et al. (2019). The Gram–Schmidt Walk algorithm produces a random assignment vector $\boldsymbol{z} \in \{\pm 1\}^n$ so that the (random) difference between the within-group sums of the augmented vectors concentrates with high probability around zero, $\boldsymbol{B}\boldsymbol{z} = \sum_{i \in Z^+} \boldsymbol{b}_i - \sum_{i \in Z^-} \boldsymbol{b}_i \approx 0$. We defer the technical discussion of the exact nature of the concentration statements to later sections.

By balancing the augmented covariate vectors, the GSW-Design balances both the original raw covariate vectors as well as the unit-unique basis vectors. The design parameter ϕ determines to what extent the augmented covariate vectors resemble either the raw covariate vectors or the orthogonal basis vectors, and thus to what extent each of these sets of vectors are balanced. This is the key way in which GSW-Design navigates the robustness-balance trade-off. The algorithm for sampling from

the GSW-Design is given formally below as Algorithm 1.

Algorithm 1: Gram–Schmidt Walk

Initialize a vector of *fractional assignments* $\boldsymbol{z}_1 \leftarrow (0, 0, \ldots, 0)$.
Initialize an index $t \leftarrow 1$.
Select an initial *pivot* unit p uniformly at random from $[n]$.
while $\boldsymbol{z}_t \notin \{\pm 1\}^n$ **do**
  Create the set $\mathcal{A} \leftarrow \{i \in [n] : |\boldsymbol{z}_t(i)| < 1\}$.
  If $p \notin \mathcal{A}$, select a new pivot p from $\mathcal{A}$ uniformly at random.
  Compute a *step direction* as

$$\begin{aligned} \boldsymbol{u}_t \leftarrow \underset{\boldsymbol{u}}{\operatorname{argmin}} \quad & \|\boldsymbol{B}\boldsymbol{u}\|^2 \\ \text{subject to} \quad & u(i) = 0 \text{ for all } i \notin \mathcal{A} \\ & u(p) = 1 \end{aligned}$$

  Set $\delta^+ \leftarrow |\max \Delta|$ and $\delta^- \leftarrow |\min \Delta|$ where $\Delta = \{\delta \in \mathbb{R} : \boldsymbol{z}_t + \delta \boldsymbol{u}_t \in [-1, 1]^n\}$.
  Select a *step size* at random according to

$$\delta_t \leftarrow \begin{cases} \delta^+ & \text{with probability } \delta^-/(\delta^+ + \delta^-), \\ -\delta^- & \text{with probability } \delta^+/(\delta^+ + \delta^-). \end{cases}$$

  Update the fractional assignments: $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t + \delta_t \boldsymbol{u}_t$.
  Increment the index: $t \leftarrow t + 1$.
return *assignment vector* $\boldsymbol{z}_t \in \{\pm 1\}^n$.

The goal of the algorithm is to make the (random) discrepancy vector of the augmented covariates $\boldsymbol{B}\boldsymbol{z}$ concentrate around 0 with high probability. The algorithm takes on this balancing problem using a relaxation. In particular, the algorithm relaxes the assignments from the integral values $\{\pm 1\}$ to the interval $[-1, 1]$. We refer to assignments in the interior of this interval as *fractional*. The algorithm constructs the assignments by iteratively updating a vector of fractional assignments $\boldsymbol{z}_t$.

The initial fractional assignments are zero: $\boldsymbol{z}_1 = \boldsymbol{0}$. This means that the augmented covariate vectors start out perfectly balanced, because $\boldsymbol{B}\boldsymbol{z}_1 = \boldsymbol{B}\boldsymbol{0} = \boldsymbol{0}$. The initial assignments are not acceptable, however, because they are not ± 1. The only acceptable outputs are assignments $\boldsymbol{z}_t \in \{\pm 1\}^n$. As the algorithm updates the fractional assignments, the fundamental tension is between maintaining good balance, as measured by $\boldsymbol{B}\boldsymbol{z}_t$, and making the assignments ± 1. As we move towards ± 1, balance becomes harder to maintain. The algorithm navigates this tension by updating the assignments in a direction that does not increase the imbalances too much, while ensuring that the update is large enough to be a sizable step towards integrality.

The fractional assignments are updated by

$$\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t + \delta_t \boldsymbol{u}_t.$$

The update $\delta_t \boldsymbol{u}_t$ is comprised of a step size δ_t and a step direction $\boldsymbol{u}_t$. The algorithm selects the step direction to minimize the imbalance of the update as measured by the magnitude of the balance of the augmented covariate vectors:

$$\|\boldsymbol{B}\boldsymbol{u}_t\|^2 = \left\| \sum_{i=1}^{n} u_t(i) \boldsymbol{b}_i \right\|^2 .$$

As the update is additive, we have $\boldsymbol{B}\boldsymbol{z}_{t+1} = \boldsymbol{B}\boldsymbol{z}_t + \delta_t \boldsymbol{B}\boldsymbol{u}_t$. In this way, making $\|\boldsymbol{B}\boldsymbol{u}_t\|^2$ small helps keep $\|\boldsymbol{B}\boldsymbol{z}_{t+1}\|^2$ small.

The update direction is selected under two constraints. The first is that the coordinates corresponding to units that already have ± 1 assignments are zero. That is, we impose $u(i) = 0$ for all $i \notin \mathcal{A}$. The purpose is to ensure that these units maintain their ± 1 assignments. The second constraint is that the coordinate for one unit $p \in \mathcal{A}$, which we call the pivot, is one: $u(p) = 1$. The pivot fills two purposes: the first purpose is to avoid the trivial solution $\boldsymbol{u}_t = \boldsymbol{0}$. The second purpose is to avoid compounding imbalances in the updates, which we discuss more in Section 2.3.3.

With the step direction in hand, the algorithm randomly selects the step size δ_t to be one of two candidate values: δ^+ and δ^-. The candidate values, one positive and one negative, are the largest scaling factors δ_t such that the updated assignment vector $\boldsymbol{z}_t + \delta_t \boldsymbol{u}_t$ is in the cube $[-1, 1]^n$. This ensures that the updated assignments are valid fractional assignments. It also ensures that at least one unit with an assignment in the interior of the interval is assigned ± 1 at each iteration. The procedure is repeated until a ± 1 assignment vector is reached.

Our implementation of the GSW-Design differs from the original algorithm of Bansal et al. (2019) only in the choice of pivot unit: we select the pivot uniformly at random whereas Bansal et al. (2019) deterministically selects the pivot to be the largest unit. This difference is quite minor and only plays a role in variance estimation, as discussed in Section 2.6.2. A comprehensive comparison between the two implementations is discussed in Appendix A.1.5.

Figure 2.1 provides an illustration of the algorithm. Panel A depicts the fractional assignments as an update iteration starts in the third step. Panel B depicts the selected update direction. This direction depends on the augmented covariates, which are not illustrated in the figure. Panels C and D show the two possible updates given by the two candidate step sizes. Panel E depicts the randomly updated assignment vector at the end of the iteration.

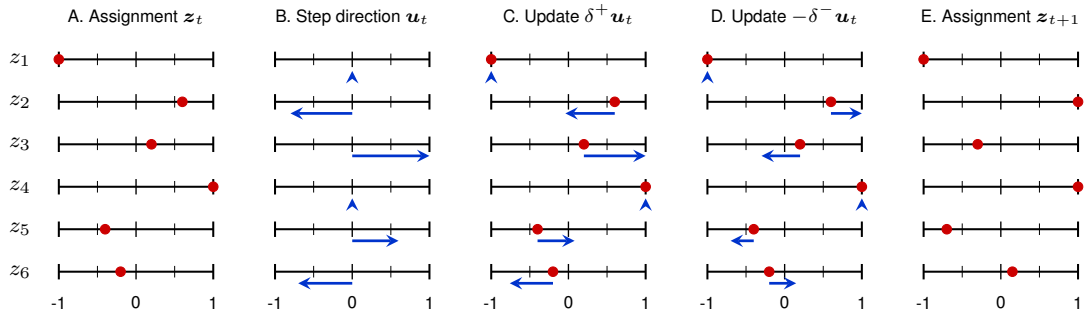

Figure 2.1: Illustration of one iteration of the Gram–Schmidt Walk design.

2.3.1 Individual treatment probabilities

We now establish that under the GSW-Design, each unit is equally likely to receive either treatment assignment. The key insight is that the (random) iterates of the fractional assignment vector form a martingale, as shown in the following Lemma:

Lemma 2.7. *The sequence of fractional assignments $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$ forms a martingale.*

Proof. Recall that the fractional assignments are updated as $\boldsymbol{z}_{t+1} = \boldsymbol{z}_t + \delta_t \boldsymbol{u}_t$. Consider the conditional expectation of the assignments updated at iteration t:

$$\mathbb{E}[\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t] = \boldsymbol{z}_t + \mathbb{E}[\delta_t \boldsymbol{u}_t \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t].$$

By the law of iterated expectations,

$$\mathbb{E}[\delta_t \boldsymbol{u}_t \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t] = \mathbb{E}\big[\mathbb{E}[\delta_t \mid \delta_t^+, \delta_t^-]\boldsymbol{u}_t \;\big|\; \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t\big],$$

because δ_t is conditionally independent of $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_t, \boldsymbol{u}_t)$ given (δ_t^+, δ_t^-). The step size δ_t takes the values δ_t^+ and δ_t^- with probabilities inversely proportional to their magnitudes, so

$$\mathbb{E}[\delta_t \mid \delta_t^+, \delta_t^-] = \delta_t^+ \left(\frac{\delta_t^-}{\delta_t^+ + \delta_t^-}\right) - \delta_t^- \left(\frac{\delta_t^+}{\delta_t^+ + \delta_t^-}\right) = 0.$$

It follows that the expected update is zero: $\mathbb{E}[\delta_t \boldsymbol{u}_t \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t] = \boldsymbol{0}$. □

The martingale property implies that the expectation of the assignments sampled from the design is zero: $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{z}_1 = \boldsymbol{0}$. This yields the following corollary, which follows from $\mathbb{E}[z_i] = \Pr(z_i = 1) - \Pr(z_i = -1) = 0$ for all units.

Corollary 2.8. *Under the* GSW-Design, $\Pr(z_i = 1) = 1/2$ *for all* $i \in [n]$.

Because the marginal treatment probabilities are bounded away from 0, the following corollary follows from Proposition 1.2.

Corollary 2.9. *The Horvitz–Thompson estimator is unbiased for the average treatment effect under the* GSW-Design.

The relation $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{z}_1$ holds for any initial fractional assignments, which provides control over the first moment of the assignment vector. We use this insight to extend the design to non-uniform assignment probabilities in Appendix A.1.6.

2.3.2 Efficient $\mathcal{O}(n^2 d)$ implementation

The structure of the augmented covariates allow us to construct a customized implementation of the Gram–Schmidt Walk algorithm that is considerably faster than a general implementation. Appendix A.2 describes this implementation and proves its computational properties. The results are summarized here.

Lemma 2.10. *The Gram–Schmidt Walk terminates after at most n iterations.*

Proof. The step direction is selected under the condition that the coordinates of units with integral assignments are zero. As a consequence, once a unit is assigned an integral assignment, it keeps that assignment. Furthermore, the candidate step sizes are selected so that at least one fractional assignment is updated to be integral at every iteration. The implication is that the number of units with integral assignments grows by at least one per iteration. □

At each iteration, the most computationally intensive operation is the computation of the step direction, $\boldsymbol{u}_t$. This is a least squares problem, so the solution can be obtained exactly by solving a system of linear equations. The number of equations at each iteration is $\mathcal{O}(n)$ so that the linear system may be solved using $\mathcal{O}(n^3)$ arithmetic operations. Thus, a naive implementation of GSW-Design requires $\mathcal{O}(n^4)$ arithmetic operations to sample an assignment vector.

We obtain a faster implementation which exploits the structure of the augmented covariates and the repeated linear system solves. In particular, an application of the Woodbury matrix inversion identity allows us to reduce the least squares problem to a linear system with d equations followed by a matrix–vector multiplication. In addition, we maintain a matrix factorization of this smaller linear system for faster repeated solves. The matrix–vector multiplication requires $\mathcal{O}(nd)$ arithmetic operations and the small linear system solve requires $\mathcal{O}(d^2)$ arithmetic operations with the factorization. Together, these two techniques allow us to solve the least squares problem at each iteration using $\mathcal{O}(nd)$ arithmetic operations. These improvements yield an improved implementation to sample an assignment vector from the GSW-Design using only $\mathcal{O}(n^2 d)$ arithmetic operations, which is a significant improvement over the naive implementation.

Proposition 2.11. *Assignments from the Gram–Schmidt Walk design can be sampled using $\mathcal{O}(n^2 d)$ arithmetic operations and $\mathcal{O}(n + d^2)$ additional storage.*

The proposition tells us that sampling an assignment from the design requires roughly the same computational resources as computing all pairwise inner products between the covariate vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The run time of the Gram–Schmidt Walk

design is therefore on the same order as a greedy implementation of the matched pair design.

2.3.3 Solving the distributional discrepancy problem

We now demonstrate that the GSW-DESIGN solves Problem 2.5 with $C = 1$. The key technical aspect is the following theorem, which bounds the resulting covariance of the discrepancy of the augmented covariate vectors $\mathrm{Cov}(\boldsymbol{Bz})$ in the Loewner partial order.

Theorem 2.12. *Under the Gram–Schmidt Walk design, the covariance matrix of the vector of imbalances for the augmented covariates $\boldsymbol{Bz}$ is bounded in the Loewner order by the orthogonal projection onto the subspace spanned by the columns of $\boldsymbol{B}$:*

$$\mathrm{Cov}(\boldsymbol{Bz}) \preceq \boldsymbol{P} \triangleq \boldsymbol{B}\big(\boldsymbol{B}^\intercal \boldsymbol{B}\big)^{-1}\boldsymbol{B}^\intercal.$$

Sketch of proof. We will show that $\boldsymbol{v}^\intercal \mathrm{Cov}(\boldsymbol{Bz})\boldsymbol{v} \leq \boldsymbol{v}^\intercal \boldsymbol{P}\boldsymbol{v}$ for all vectors $\boldsymbol{v} \in \mathbb{R}^{n+d}$. In Appendix A.1.3, we derive an expression for $\mathrm{Cov}(\boldsymbol{z})$ in terms of the step directions and sizes used by the algorithm in Section 4.6. This allows us to write the quadratic form as

$$\boldsymbol{v}^\intercal \mathrm{Cov}(\boldsymbol{Bz})\boldsymbol{v} = \boldsymbol{v}^\intercal \,\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \boldsymbol{B}\boldsymbol{u}_t\boldsymbol{u}_t^\intercal \boldsymbol{B}^\intercal\right]\boldsymbol{v} = \mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right],$$

where T is the final iteration of the algorithm. Note that T is random.

The first part of the proof is to rearrange the terms of this sum. To do so, we define a pivot phase S_i as the set of iterations t for which unit i was the pivot. A unit's pivot phase is random and it may be the empty set if the unit was assigned a ± 1 without being chosen as the pivot. We can now write

$$\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\sum_{t\in S_i} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right].$$

In the appendix, we show that the expected sum of the squared step sizes within a pivot phase is bounded by one. This is a consequence of the fact that the same unit is kept as pivot until it is assigned a value in ± 1. Together with the fact that each column of $\boldsymbol{B}$ has norm of at most one, this allows us to bound the contribution of each pivot phase to the overall quadratic form as

$$\mathbb{E}\left[\sum_{t\in S_i} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right] \leq \mathbb{E}\big[\boldsymbol{v}^\intercal \boldsymbol{P}_i \boldsymbol{v}\big],$$

where $\boldsymbol{P}_i$ denotes the projection onto a subspace that contains the updates $\boldsymbol{B}\boldsymbol{u}_t$ generated in the pivot phase S_i.

Bansal et al. (2019) show that the updates $\boldsymbol{B}\boldsymbol{u}_t$ and $\boldsymbol{B}\boldsymbol{u}_s$ are orthogonal if the iterations t and s are in different pivot phases. In the appendix, we extend this result to show that the subspaces corresponding to different pivot phases are orthogonal and their union is the column space of $\boldsymbol{B}$, so that $\sum_{i=1}^{n} \boldsymbol{P}_i = \boldsymbol{P}$ with probability one. We conclude that

$$\sum_{i=1}^{n} \mathbb{E}\big[\boldsymbol{v}^\mathsf{T} \boldsymbol{P}_i \boldsymbol{v}\big] = \boldsymbol{v}^\mathsf{T}\, \mathbb{E}\left[\sum_{i=1}^{n} \boldsymbol{P}_i\right] \boldsymbol{v} = \boldsymbol{v}^\mathsf{T} \boldsymbol{P} \boldsymbol{v}. \qquad \square$$

We provide a detailed proof of the theorem in Appendix A.1.3. Our proof interprets the procedure as implicitly constructing a random basis for the column space of $\boldsymbol{B}$. This reveals the connection between the Gram–Schmidt Walk and its namesake, the Gram–Schmidt orthogonalization procedure.

Theorem 2.12 therefore demonstrates that the design, as intended, balances the augmented covariates. The projection matrix $\boldsymbol{P}$ is small: it has at most n eigenvalues that are one and d eigenvalues that are zero. Another way to see that the augmented covariates are well-balanced is to consider the variance of linear functions of the augmented covariates. For every $\boldsymbol{v} \in \mathbb{R}^{n+d}$,

$$\operatorname{Var}\Big(\sum_{i \in Z^+} \langle \boldsymbol{v}, \boldsymbol{b}_i \rangle - \sum_{i \in Z^-} \langle \boldsymbol{v}, \boldsymbol{b}_i \rangle\Big) = \boldsymbol{v}^\mathsf{T} \operatorname{Cov}(\boldsymbol{B}\boldsymbol{z}) \boldsymbol{v} \leq \boldsymbol{v}^\mathsf{T} \boldsymbol{P} \boldsymbol{v} \leq \|\boldsymbol{v}\|^2,$$

where the inequalities follow from Theorem 2.12 and the fact that projection operators are contractive.

We now show that the Gram–Schmidt Walk algorithm solves Problem 2.5 with $C = 1$. We emphasize that in the later sections of the chapter, we provide an improved analysis of covariate balancing and mean squared error bounds. The main idea is to extract the principal submatrices of the matrix inequality given by Theorem 2.12.

Corollary 2.13. *The* GSW-Design *satisfies the requirement of Problem 2.5 with* $C = 1$. *In particular, the random assignment vector* $\boldsymbol{z}$ *may be sampled used* $\mathcal{O}(n^2 d)$ *arithmetic operations and satisfies* $\Pr(z_i = 1) = 1/2$ *for all* $i \in [n]$ *and the operator norm bounds*

$$\|\operatorname{Cov}(\boldsymbol{z})\| \leq \frac{1}{\phi} \quad \textit{and} \quad \frac{1}{\xi^2} \|\operatorname{Cov}(\boldsymbol{X}^\mathsf{T} \boldsymbol{z})\| \leq \frac{1}{1-\phi} \ .$$

Proof. The runtime is established in Proposition 2.11 and the uniform marginal probability of treatment assignment is established in Corollary 2.8. Thus, it remains to establish the operator norm bounds.

To this end, recall that all projection matrices are less than the identity matrix in the Loewner order. Thus, Theorem 2.12 implies that $\operatorname{Cov}(\boldsymbol{B}\boldsymbol{z}) \preceq \boldsymbol{P} \preceq \boldsymbol{I}$. Observe that the covariance matrix of $\operatorname{Cov}(\boldsymbol{B}\boldsymbol{z})$ can be written in block form as

$$\operatorname{Cov}(\boldsymbol{B}\boldsymbol{z}) = \begin{bmatrix} \phi \operatorname{Cov}(\boldsymbol{z}) & \xi^{-1}\sqrt{\phi(1-\phi)}\operatorname{Cov}(\boldsymbol{X}^\mathsf{T}\boldsymbol{z}, \boldsymbol{z})^\mathsf{T} \\ \xi^{-1}\sqrt{\phi(1-\phi)}\operatorname{Cov}(\boldsymbol{X}^\mathsf{T}\boldsymbol{z}, \boldsymbol{z}) & \xi^{-2}(1-\phi)\operatorname{Cov}(\boldsymbol{X}^\mathsf{T}\boldsymbol{z}) \end{bmatrix}.$$

By extracting the upper left and lower right blocks in the matrix inequality $\mathrm{Cov}(\boldsymbol{B}\boldsymbol{z}) \preceq \boldsymbol{I}$ and rearranging terms, we have that

$$\mathrm{Cov}(\boldsymbol{z}) \preceq \frac{1}{\phi}\boldsymbol{I} \quad \text{and} \quad \frac{1}{\xi^2}\mathrm{Cov}(\boldsymbol{X}^\mathsf{T}\boldsymbol{z}) \preceq \frac{1}{1-\phi}\boldsymbol{I} \ .$$

The claim is established by taking the operator norm of both sides. □

This demonstrates the way in which the GSW-Design navigates the balance-robustness trade-off. In the later sections, we more closely analyze various aspects of the GSW-Design, including the mean-squared error of the Horvitz–Thompson estimator, the covariate balancing properties of the design, and the tail behavior of the Horvitz–Thompson estimator.

2.4 Analysis of the Mean Squared Error

In this section, we present a more refined analysis of the mean squared error of the Horvitz–Thompson estimator under the GSW-Design. This analysis demonstrates that GSW-Design acts in a way which we call *regression by design*, as discussed in Section 2.2.4. Finally, we discuss several methods for choosing the design parameter $\phi \in [0, 1]$ based on finite sample and asymptotic analyses.

2.4.1 Refined bound on the mean squared error

Because GSW-Design satisfies the requirements of Problem 2.5, the mean squared error of the Horvitz–Thompson estimator under GSW-Design is upper bounded by Theorem 2.6. However, we can obtain a tighter bound on the mean squared error by more closely analyzing the crossing terms. The following theorem, which bounds the mean squared error, follows from a more careful application of Theorem 2.12. Its proof appears in Appendix A.3.1

Theorem 2.14. *The mean squared error under the* GSW-Design *is at most the minimum of the loss function of an implicit ridge regression of the sum of the potential outcome vectors $\boldsymbol{\mu} = (\boldsymbol{a} + \boldsymbol{b})$ on the covariates:*

$$\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] \leq \frac{L}{n} \qquad \textit{where} \qquad L = \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left[\frac{1}{\phi n} \big\| \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta} \big\|^2 + \frac{\xi^2}{(1-\phi)n} \big\| \boldsymbol{\beta} \big\|^2 \right].$$

Like the mean squared error bound of Theorem 2.6—which holds for any design satisfying the distributional discrepancy Problem 2.5—the mean squared error bound of Theorem 2.14 also has the interpretation of regression-by-design. In particular, the mean squared error depends on the design parameter ϕ and the degree to which the covariates are predictive of the potential outcomes. However, the improvement in Theorem 2.14 over Theorem 2.6 is the dependence on the design parameter ϕ. In

particular, the dependence is improved from ϕ^{-2} and $(1-\phi)^{-2}$ to ϕ^{-1} and $(1-\phi)^{-1}$. Concretely, this results in a factor 2 decrease in the bound of the mean-squared error when $\phi = 1/2$. This improvement is obtained by the balancing guarantees of the augmented covariates, which allows us to more carefully analyze crossing terms.

2.4.2 Choosing the design parameter ϕ

In this section, we discuss several considerations when choosing the design parameter. The first observation is the following corollary, which described conditions under which the experimenter should set $\phi < 1$.

Corollary 2.15. *If the scaled sum of cross-moments between covariates and potential outcomes is greater than the second moment of potential outcomes, $\xi^{-2}\|\boldsymbol{X}^\intercal\boldsymbol{\mu}\|^2 > \|\boldsymbol{\mu}\|^2$, then the design parameter ϕ that minimizes the mean squared error is less than one.*

The corollary provides precise conditions for when it is beneficial to deviate from the fully randomized design we get when $\phi = 1$. The cross-moments capture the predictiveness of the covariates. To see this, consider when the covariates and potential outcomes are demeaned, in which case $n^{-2}\|\boldsymbol{X}^\intercal\boldsymbol{\mu}\|^2$ is the sum of squared covariances between covariates and potential outcomes, and $n^{-1}\|\boldsymbol{\mu}\|^2$ is the variance of the potential outcomes. Therefore, the left-hand side becomes larger as the covariates become more predictive.

Note that $\|\boldsymbol{X}^\intercal\boldsymbol{\mu}\|^2$ tends to grow at an n^2-rate if the covariates remain predictive asymptotically, while $\|\boldsymbol{\mu}\|^2$ tends to grow at an n-rate. Therefore, the left-hand side is generally much larger than the right-hand side in large samples even if the covariates are only weakly predictive. The factor ξ^2 captures the scaling of the covariates and the presence of outliers. For well-behaved sets of covariate vectors without extreme outliers (e.g. covariate vectors sampled from a subgaussian distribution), a reasonable growth rate of ξ^2 is $d\log(n)$, meaning that the scaling will generally not be consequential. This tells us that it is almost always beneficial to at least partially balance the covariates in large samples, so we should set $\phi < 1$. One exception is experiments with many unpredictive covariates, where $\phi = 1$ often will be optimal.

The following corollary bounds the mean squared error of the Horvitz–Thompson estimator under the GSW-Design in large samples when the design parameter $\phi < 1$ is fixed.

Corollary 2.16. *Let $\boldsymbol{\beta}_{\text{LS}} \in \arg\min_{\boldsymbol{\beta}}\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|$ be the best least squares linear approximator of the potential outcomes with smallest norm, and let $\boldsymbol{\varepsilon} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}$ be the errors of those approximations. Fix a design parameter $\phi < 1$. If $\|\boldsymbol{\beta}_{\text{LS}}\|^2 = o\big(\xi^{-2}n\big)$, then the normalized mean squared error under the* GSW-Design *is asymptotically upper bounded by*

$$\limsup_{n\to\infty}\left[n\,\mathbb{E}\big[(\widehat{\tau}-\tau)^2\big] - \frac{1}{\phi n}\|\boldsymbol{\varepsilon}\|^2\right] \leq 0.$$

The corollary characterizes the large sample behavior of the design. The condition $\|\boldsymbol{\beta}_{\text{LS}}\|^2 = o\big(\xi^{-2}n\big)$ states that the linear coefficients do not diverge at a too fast rate asymptotically. In Appendix A.3.2, we show that this condition is satisfied if the second moment of the potential outcomes $\|\boldsymbol{\mu}\|^2/n$ stays bounded and the maximum row norm ξ is asymptotically dominated by the smallest, non-zero singular value of $\boldsymbol{X}$. This is the case, for example, when the covariates are of fixed dimension and not nearly multicollinear, so that $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ is invertible.

The asymptotic analysis of Corollary 2.16 suggests a simple heuristic for selecting the design parameter: set $\phi \geq \|\boldsymbol{\varepsilon}\|^2/\|\boldsymbol{\mu}\|^2$, which is equal to $1 - R^2$, where R^2 is the coefficient of determination. Recall that the normalized mean squared error under the fully randomized design, which is minimax optimal, is $\|\boldsymbol{\mu}\|^2/n$. Thus, if we set $\phi \geq \|\boldsymbol{\varepsilon}\|^2/\|\boldsymbol{\mu}\|^2$, the error under the Gram–Schmidt Walk design is asymptotically no worse than the minimax design. For example, if the covariates are only somewhat predictive, so that $R^2 = 0.1$, then the heuristic stipulates that we set ϕ to a value larger than $1 - 0.1 = 0.9$. Of course, the R^2 value cannot be exactly known before running the experiment, but it provides a useful heuristic that can leverage the experimenter's prior substantive knowledge.

The following corollary demonstrates that an improved mean squared error is achievable if we let the design parameter vary in the asymptotic sequence.

Corollary 2.17. *Under the conditions of Corollary 2.16, the normalized mean squared error under the* GSW-Design *with the adaptive parameter choice of* $\phi = \big(1 + \xi\|\boldsymbol{\beta}_{\text{LS}}\|/\|\boldsymbol{\varepsilon}\|\big)^{-1}$ *is asymptotically upper bounded by*

$$\limsup_{n\to\infty}\left[n\,\mathbb{E}\big[(\widehat{\tau}-\tau)^2\big] - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\right] \leq 0.$$

Note that a normalized mean squared error of $\|\boldsymbol{\varepsilon}\|^2/n$ would be attainable if we somehow had access to all potential outcomes before the experiment started, so we could calculate $\boldsymbol{\beta}_{\text{LS}}$, and then used the residuals $\boldsymbol{\varepsilon}$ from this regression as outcomes in the experiment. In this sense, $\|\boldsymbol{\varepsilon}\|^2/n$ marks the lowest normalized mean squared error achievable by balancing linear functions. Corollary 2.17 shows that, in this particular asymptotic regime, we can attain this lower limit when if we carefully allow the parameter ϕ to approach 1 with the sample size. It is important to note that we attain this lower limit by letting the design parameter approach one, but we cannot set it exactly to one. If we were to set $\phi = 1$, we would get the fully randomized design, and the normalized mean squared error would be $\|\boldsymbol{\mu}\|^2/n$. Because the design parameter varies with the asymptotic sequence in Corollary 2.17, experimenters may find this result less helpful in setting the design parameter.

2.5 Analysis of Covariate Balancing

In this section, we investigate the covariate balancing properties of the GSW-Design. First, we obtain a more refined analysis of the covariate balance than what is guaranteed by Problem 2.5. Then, we use an existing hardness result to show that improving the covariate balance by even a constant factor is computationally intractable.

2.5.1 Refined bound on covariate balance

We now present a more refined analysis of the covariate balancing properties of the GSW-Design than the operator norm bound in Problem 2.5. We begin by presenting a matrix bound on the covariance matrix of the discrepancy vector of covariates.

Proposition 2.18. *Under the* GSW-Design*, the covariance matrix of $\boldsymbol{X}^\intercal \boldsymbol{z}$ is bounded in the Loewner order by*

$$\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z}) \preceq \Big(\phi(\boldsymbol{X}^\intercal \boldsymbol{X})^\dagger + (1-\phi)(\xi^2 \boldsymbol{\Pi})^\dagger\Big)^\dagger,$$

where $\boldsymbol{\Pi}$ is the orthogonal projection onto the rows of the covariate matrix $\boldsymbol{X}$ and $\boldsymbol{A}^\dagger$ denotes the pseudo-inverse of $\boldsymbol{A}$.

The matrix in the upper bound is the weighted harmonic mean of two d-by-d matrices: the Gram matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$ and the scaled projection matrix $\xi^2 \boldsymbol{\Pi}$. When $\phi = 1$, the bound is the Gram matrix, which is the value the covariate matrix takes when the assignments are pair-wise indpendent. When $\phi = 0$, the bound is $\xi^2 \boldsymbol{\Pi}$, which is a scaled version of the projection onto the span of the covariate vectors. When the covariate vectors span the entire vector space, $\boldsymbol{\Pi}$ is the identity matrix; otherwise, we may interpret $\boldsymbol{\Pi}$ as being the identity matrix on the subspace containing the data. Intermediate values interpolate between the two extremes.

The matrix bound in Proposition 2.18 yields a bound on the variance of the difference between the within-group sums of any linear functions of the covariate vectors. In particular, applying the definition of the Loewner order and evaluating the quadratic form, we have that for any linear function $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\mathbb{E}\left[\Big(\sum_{i \in Z^+} \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle - \sum_{i \in Z^-} \langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle\Big)^2\right] \leq \boldsymbol{\theta}^\intercal \Big(\phi(\boldsymbol{X}^\intercal \boldsymbol{X})^\dagger + (1-\phi)(\xi^2 \boldsymbol{\Pi})^\dagger\Big)^\dagger \boldsymbol{\theta} \ .$$

When $\boldsymbol{\theta}$ is a basis vector, then the inequality above bounds the discrepancy of a single covariate between the two groups. The inequality may be hard to interpret for a general linear function, but experimenters may use the quadratic form on the right hand side to investigate an imbalances in the covariates before running the experiment. In any case, we may use the operator norm bound on the quadratic form

on the right hand side to obtain a worst-case bound over all linear functions:

$$\mathbb{E}\Big[\Big(\sum_{i\in Z^+}\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle - \sum_{i\in Z^-}\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\Big)^2\Big] \leq \frac{\|\boldsymbol{\theta}\|^2}{\phi\lambda_G^{-1} + (1-\phi)\xi^{-2}}\ , \tag{2.1}$$

where λ_G is the largest eigenvalue of the Gram matrix $\boldsymbol{X}^\intercal\boldsymbol{X}$. This bound mirrors the matrix bound in Proposition 2.18, in that it is a weighted harmonic mean between λ_G and ξ^2. At the extremes, when ϕ is either one or zero, the bound is λ_G and ξ^2, respectively. Intermediate values of ϕ interpolate between the two end points.

The interpolation is monotone: the bound decreases with ϕ. This is because $\lambda_G \geq \xi^2$. This indicates that the imbalance for the worst-case linear function tends to decrease as the parameter approaches zero. Moreover, (2.1) shows that the magnitude of λ_G relative to ξ determines the slope of the decrease. The eigenvalue λ_G is typically considerably larger than the norm ξ, so the imbalance tends to decrease quickly with ϕ. To see this, let $k \in [n]$ be such that $\|\boldsymbol{x}_k\| = \xi = \max_{i\in[n]}\|\boldsymbol{x}_i\|$, and observe that

$$\lambda_G = \max_{\|\boldsymbol{\theta}\|\leq 1}\sum_{i=1}^n \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle^2 \geq \max_{\|\boldsymbol{\theta}\|\leq 1}\langle \boldsymbol{x}_k, \boldsymbol{\theta}\rangle^2 = \|\boldsymbol{x}_k\|^2 = \xi^2.$$

The gap introduced by the inequality is large as long as there is not a unit whose covariate vector has disproportionately large norm and is nearly orthogonal to the vectors of the other units. The fewer outliers there are, the larger λ_G will be relative to ξ^2, and the more balance can be achieved.

We remark that no design can improve upon Proposition 2.18 without imposing structural restrictions on the covariates. In particular, the scaling term ξ^2 cannot be improved for general covariate vectors. To see this, consider a set of covariate vectors where one $\boldsymbol{x}_k$ is orthogonal to the remaining covariate vectors and has the largest norm, $\xi = \|\boldsymbol{x}_k\|$. In this case, by choosing the linear function $\boldsymbol{\theta} = \boldsymbol{x}_k$, we have that

$$\begin{aligned}
\mathbb{E}\Big[\Big(\sum_{i\in Z^+}\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle - \sum_{i\in Z^-}\langle \boldsymbol{\theta}, \boldsymbol{x}_i\rangle\Big)^2\Big] &= \mathbb{E}\Big[\Big(\sum_{i\in Z^+}\langle \boldsymbol{x}_k, \boldsymbol{x}_i\rangle - \sum_{i\in Z^-}\langle \boldsymbol{x}_k, \boldsymbol{x}_i\rangle\Big)^2\Big] && \text{(choice of } \boldsymbol{\theta}\text{)} \\
&= \mathbb{E}\Big[z_k^2\langle \boldsymbol{x}_k, \boldsymbol{x}_k\rangle^2\Big] && \text{(orthogonality)} \\
&= \|\boldsymbol{x}_k\|^4 = \xi^2\cdot\|\boldsymbol{\theta}\|^2\ , && \text{(choice of } \boldsymbol{\theta}\text{)}
\end{aligned}$$

which demonstrates that (2.1) is tight when $\phi = 0$. In this example, the covariate vector $\boldsymbol{x}_k$ may be considered an outlier. Generally speaking, better covariate balancing guarantees will not be possible in the presence of outliers.

2.5.2 Computational barriers to improved covariate balance

In this section, we demonstrate that achieving more covariate balance than that which is guaranteed by Gram–Schmidt Walk design with $\phi = 0$ is computationally

intractable. Charikar et al. (2011) prove that, given an n-by-n matrix $\boldsymbol{X}$ with ± 1 entries, it is NP-hard to determine whether

$$\min_{\boldsymbol{z} \in \{\pm 1\}^n} \|\boldsymbol{X}^\mathsf{T} \boldsymbol{z}\|^2 \geq c\, n^2 \quad \text{or} \quad \min_{\boldsymbol{z} \in \{\pm 1\}^n} \|\boldsymbol{X}^\mathsf{T} \boldsymbol{z}\|^2 = 0,$$

where $c > 0$ is universal, but presently unspecified, constant. We compare this hardness result to the covariate balance guarantees we prove for the Gram–Schmidt Walk design with $\phi = 0$. The covariate balance guarantees of Proposition 2.18 imply that in this case,

$$\mathbb{E}\big[\|\boldsymbol{X}^\mathsf{T} \boldsymbol{z}\|^2\big] = \operatorname{tr}(\operatorname{Cov}(\boldsymbol{X}^\mathsf{T} z)) \leq \xi^2 \operatorname{tr}(\boldsymbol{\Pi}) \leq n^2,$$

where the third inequality follows by properties of projection matrices and that $\boldsymbol{X}$ $\boldsymbol{X}$ has ± 1 entries, so $\xi^2 = n$. Thus, improving the covariate balance by even a constant factor pushes up against the boundary of computational tractability. This demonstrates that no computationally feasible design can provide a significantly better guarantee on expected covariate balance without assumptions on the structure of the covariates.

2.6 Analysis of Tail Behavior

The previous sections examined the precision of the Horvitz–Thompson under the GSW-Design in the mean square sense. In this section, we extend the investigation of precision to tail behavior by deriving a subgaussian tail bound. This provides an alternative and often sharper description of the properties of the design. Finally, we discuss how to use this tail inequality to construct confidence intervals.

2.6.1 Subgaussian tail bounds

Bansal et al. (2019) used the martingale inequality of Freedman (1975) to show that the Gram–Schmidt Walk algorithm produces assignments such that $\boldsymbol{B}\boldsymbol{z}$ is a subgaussian random vector with variance parameter $\sigma^2 \leq 40$. This result allows us to investigate the behavior of the design in terms tail probabilities. The concern is that tail bounds based on $\sigma^2 = 40$ will generally be too loose to be useful in a statistical context. Unless we are interested in the extreme ends of the tails, Chebyshev's inequality based on the mean squared error results in Section 2.4 will be more informative.

An important contribution of this chapter is to strengthen the analysis of the tail behavior of the Gram–Schmidt Walk algorithm. We develop a new proof technique for establishing martingale concentration, thus obtaining a tight upper bound on the subguassian parameter.

Theorem 2.19. *Under the Gram–Schmidt Walk design, the vector $\boldsymbol{Bz}$ is subgaussian with variance parameter $\sigma^2 = 1$:*

$$\mathbb{E}\big[\exp\big(\langle \boldsymbol{Bz}, \boldsymbol{v}\rangle\big)\big] \leq \exp\big(\|\boldsymbol{v}\|^2/2\big) \qquad \textit{for all} \qquad \boldsymbol{v} \in \mathbb{R}^{n+d}.$$

Sketch of Proof. Recall the projection matrix $\boldsymbol{P} = \boldsymbol{B}\big(\boldsymbol{B}^\intercal \boldsymbol{B}\big)^{-1}\boldsymbol{B}^\intercal$ from Theorem 2.12. Because projection is a contractive operator, we have

$$\exp\big(\|\boldsymbol{Pv}\|^2/2\big) \leq \exp\big(\|\boldsymbol{v}\|^2/2\big) \qquad \text{for all} \qquad \boldsymbol{v} \in \mathbb{R}^{n+d}.$$

Therefore, to prove the theorem, it suffices to show that

$$\mathbb{E}\left[\exp\Big(\langle \boldsymbol{Bz}, \boldsymbol{v}\rangle - \|\boldsymbol{Pv}\|^2/2\Big)\right] \leq 1.$$

Following the proof of Theorem 2.12, we decompose the assignment vector into its fractional updates and then group them according to pivot phases,

$$\langle \boldsymbol{Bz}, \boldsymbol{v}\rangle = \sum_{t=1}^{T} \delta_t \langle \boldsymbol{Bu}_t, \boldsymbol{v}\rangle = \sum_{i=1}^{n} \sum_{t \in S_i} \delta_t \langle \boldsymbol{Bu}_t, \boldsymbol{v}\rangle.$$

Similarly, we decompose the projection $\boldsymbol{P}$ into the mutually orthogonal projections given by each pivot phase:

$$\|\boldsymbol{Pv}\|^2 = \left\|\sum_{i=1}^{n} \boldsymbol{P}_i \boldsymbol{v}\right\|^2 = \sum_{i=1}^{n} \|\boldsymbol{P}_i \boldsymbol{v}\|^2,$$

where, as in the proof of Theorem 2.12, $\boldsymbol{P}_i$ denotes the projection matrix onto the subspace corresponding to pivot phase i that contains the updates $\{\boldsymbol{Bu}_t : t \in S_i\}$.

We consider the difference D_i between the two decompositions separately for each potential pivot unit i:

$$D_i = \sum_{t \in S_i} \delta_t \langle \boldsymbol{Bu}_t, \boldsymbol{v}\rangle - \|\boldsymbol{P}_i \boldsymbol{v}\|^2/2.$$

This allows us to write

$$\mathbb{E}\left[\exp\Big(\langle \boldsymbol{Bz}, \boldsymbol{v}\rangle - \|\boldsymbol{Pv}\|^2/2\Big)\right] = \mathbb{E}\left[\exp\left(\sum_{i=1}^{n} D_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n} \exp(D_i)\right].$$

If a unit is never chosen as the pivot, the corresponding pivot phase is empty and $D_i = 0$. We can therefore restrict the product to the units which at some point are pivots. For notational convenience in this proof sketch, suppose that the pivot units

are $1, 2, \ldots, r$ and they are chosen as pivots in this order. We then have

$$\mathbb{E}\left[\prod_{i=1}^{n} \exp(D_i)\right] = \mathbb{E}\left[\prod_{i=1}^{r} \exp(D_i)\right].$$

Consider a pivot unit i, where $1 \leq i \leq r$. Let Δ_i denote all random decisions made by the algorithm up to and including when i is chosen as the pivot. This includes all randomly chosen step sizes in the pivot phases $1, \ldots, i-1$, but not the step sizes in phases $i, \ldots, r$. The key part of the argument, which we prove in Appendix A.1.4, is that

$$\mathbb{E}\big[\exp(D_i) \;\big|\; \Delta_i\big] \leq 1.$$

This follows from the choice of the step sizes, the fact that a unit remains a pivot until it is assigned a ± 1, and the fact that each column of $\boldsymbol{B}$ has norm at most one.

We can now prove the inequality by backward induction. Because Δ_r includes all random decisions before unit r was selected as pivot, the quantities $D_1, \ldots, D_{r-1}$ are not random conditional on Δ_r. Using the law of iterated expectation, we can write

$$\mathbb{E}\left[\prod_{i=1}^{r} \exp(D_i)\right] = \mathbb{E}\left[\mathbb{E}\big[\exp(D_r) \;\big|\; \Delta_r\big] \prod_{i=1}^{r-1} \exp(D_i)\right] \leq \mathbb{E}\left[\prod_{i=1}^{r-1} \exp(D_i)\right].$$

The proof is completed by induction over the remaining $r-1$ pivot phases. □

The central step in the proof, which appears in Apppendix A.1.4, is bounding the conditional expectation of the exponential quantity during a pivot phase. Previous proof techniques bound this quantity through Taylor series approximations, which necessarily incur a loss in approximation and result in overly conservative subgaussian constants. In contrast, our proof analyzes the expected exponential quantity directly by carefully considering the choice of step size and another backwards induction argument. In this way, we can obtain $\sigma^2 = 1$, which is tight. This proof technique may be of independent interest for studying martingale concentration more generally.

Theorem 2.19 shows that linear functions of the augmented covariates are well concentrated. Because the augmented covariates contain the raw covariates, this implies concentration of the imbalance of any linear function of the covariates. This concentration becomes tighter as the design parameter ϕ decreases. The proof of this is analogous to the derivation of the covariate balance results in Section 2.5 using Theorem 2.12. However, in the interest of space, our focus in the rest of the section is concentration of the estimator and the construction of confidence intervals.

2.6.2 Confidence intervals

The sharpened tail bound allows us to show that the Horvitz–Thompson estimator is subgaussian as well. This yields an interval estimator for the average treatment effect. The following proposition and corollary provide the details.

Theorem 2.20. *Under the Gram–Schmidt Walk design, the mass of the tails of the sampling distribution of the Horvitz–Thompson estimator is bounded by*

$$\Pr\big(|\widehat{\tau} - \tau| \geq \gamma\big) \leq 2\exp\left(\frac{-\gamma^2 n}{2L}\right) \qquad \textit{for all} \qquad \gamma > 0.$$

Proof. We prove the bound for the upper tail. The proof for the lower tail is identical. For any $t > 0$, we have

$$\Pr\big(\widehat{\tau} - \tau \geq \gamma\big) \leq \exp(-t\gamma)\,\mathbb{E}\big[\exp\big(t(\widehat{\tau} - \tau)\big)\big].$$

This can be shown either as a consequence of Markov's inequality or from the exponential inequality $\mathbf{1}[x \geq 0] \leq \exp(tx)$. Lemma 2.1 in Section 2.1.2 shows that $\widehat{\tau} - \tau = \langle \boldsymbol{z}, \boldsymbol{\mu} \rangle / n$. The columns of $\boldsymbol{B}$ are linearly independent by construction, so we can define a vector $\boldsymbol{v} = tn^{-1}\boldsymbol{B}\big(\boldsymbol{B}^\intercal \boldsymbol{B}\big)^{-1}\boldsymbol{\mu}$. This allows us to write

$$\mathbb{E}\big[\exp\big(t(\widehat{\tau} - \tau)\big)\big] = \mathbb{E}\big[\exp\big(tn^{-1}\langle \boldsymbol{z}, \boldsymbol{\mu} \rangle\big)\big] = \mathbb{E}\big[\exp\big(\langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{v} \rangle\big)\big].$$

Theorem 2.19 upper bounds the right-hand side by $\exp(\|\boldsymbol{v}\|^2/2)$. For the current choice of $\boldsymbol{v}$, the squared norm simplifies to

$$\|\boldsymbol{v}\|^2 = \frac{t^2}{n^2}\boldsymbol{\mu}^\intercal\big(\boldsymbol{B}^\intercal\boldsymbol{B}\big)^{-1}\boldsymbol{\mu} = \frac{t^2 L}{n},$$

where the final equality follows from Lemma A.8 in Appendix A.3.1. Taken together, we obtain

$$\Pr\big(\widehat{\tau} - \tau \geq \gamma\big) \leq \exp\left(\frac{t^2 L}{2n} - t\gamma\right).$$

The proof is completed by setting $t = \gamma n / L$. □

Corollary 2.21. *The random interval centered at $\widehat{\tau}$ with radius $\gamma_\alpha = \sqrt{2\log(2/\alpha)L/n}$ is a valid $(1-\alpha)$-confidence interval:*

$$\Pr\big(\widehat{\tau} - \gamma_\alpha \leq \tau \leq \widehat{\tau} + \gamma_\alpha\big) \geq 1 - \alpha.$$

The corollary illustrates the usefulness of the sharpened tail bound in Theorem 2.19. Confidence intervals based on the tail bound in Bansal et al. (2019) would be $\sqrt{40} \approx 6.3$ times wider than the intervals in Corollary 2.21.

We emphasize here that the confidence intervals described in Corollary 2.21 cannot directly be used by an experimenter because they contain the term L/n. This term, which upper bounds the variance of the Horvitz–Thompson estimator by (Theorem 2.14), depends on all potential outcomes, half of which are unknown to the experimenter. Thus, one way to construct confidence intervals based on Corollary 2.21 is to first estimate L/n and then plug-in this estimator $\widehat{L}/n$ into the intervals described in Corollary 2.21.

Indeed, most confidence intervals proposed in the literature work in this way: first, a tail inequality or the tail of a limiting distribution is derived in which the only unknown quantity is the variance of the point estimator. Then, the variance itself must be (conservatively) estimated and this estimate is used to construct the confidence interval. Analyses of the validity of such confidence intervals are asymptotic: informally speaking, if the variance estimator is consistent or conservative, then the confidence intervals are asymptotically valid. It is a significant open problem to construct confidence intervals for average treatment effects which are valid in finite samples in the potential outcome framework considered here. Under stronger assumptions on the potential outcomes, such as constant individual treatment effect (ITE) among units, different types of confidence intervals may be obtained, e.g. Fisher type permutation tests.

Thus, in short, the quantity L/n must be estimated in order to produce confidence intervals from Corollary 2.21. This type of estimation is the focus of Chapter 3, and so we refer readers to that chapter for computational methods on estimating L. A simpler method for variance estimation is described in the working paper (Harshaw et al., 2021). Either way, the following lemma is a key insight for estimating L:

Lemma 2.22. *The second-order assignment probabilities are bounded away from zero under the* GSW-Design *for all pairs of units and all treatments:*

$$\Pr\big((z_i, z_j) = \boldsymbol{v}\big) > \frac{1}{4n} \min\Big\{\phi, \frac{\phi^2}{1-\phi}\Big\} \quad \textit{for all} \quad i \neq j \quad \textit{and all} \quad \boldsymbol{v} \in \{\pm 1\}^2.$$

In the remainder of the section, we compare confidence intervals obtained by Corollary 2.21 to confidence intervals obtained in other ways. In this discussion, we ignore the issues of variance estimation described above.

A comparison between the intervals in Corollary 2.21 and conventional intervals is intricate. One aspect is that our intervals do not rely on asymptotic approximations. This makes them particularly useful in experiments with small samples because large sample approximations may not be appropriate in such settings. However, this comes at the cost of potentially wider intervals. For example, a common approach is to approximate the distribution of the estimator with a normal distribution. Using the variance bound in Theorem 2.14, such an approach would suggest intervals with radius $\sqrt{L/n}\ \Phi^{-1}\big(1-\alpha/2\big)$ where $\Phi^{-1}\colon [0,1] \to \mathbb{R}$ is the quantile function of the standard normal deviate. Hence, for confidence levels 95% and 99%, the intervals in Corollary 2.21 would be about 1.39 and 1.26 times wider than those based on a normal approximation.

It remains an open question whether the sampling distribution of the Horvitz–Thompson estimator approaches a normal distribution under the Gram–Schmidt Walk design. Li et al. (2018) show that rerandomization does not yield estimators that are asymptotically normal. The Gram–Schmidt Walk design resembles rerandomization in some aspects, but it does not truncate the distribution of the design in the way rerandomization does. We conjecture that the Horvitz–Thompson estimator is

asymptotically normal under the GSW-Design. However, until this has been shown formally, experimenters should exercise caution when using a normal approximation even when the number of units is large.

As an illustration, consider confidence intervals based on Chebyshev's inequality. Using the variance bound in Theorem 2.14, this inequality would suggest intervals with radius $\sqrt{L/\alpha n}$. For confidence levels 95% and 99%, these intervals are about 1.6 and 3.1 times wider, respectively, than the intervals in Corollary 2.21. However, Chebyshev's inequality holds for the variance of the estimator, so we do not need to use the variance bound in Theorem 2.14. Because the bound in Theorem 2.14 can be somewhat loose, confidence intervals based on Chebyshev's inequality using the variance could be narrower than the intervals in Corollary 2.21. That is, $\mathrm{Var}(\widehat{\tau})/\alpha$ may be smaller than $2\log(2/\alpha)L/n$ because L/n is larger than $\mathrm{Var}(\widehat{\tau})$. It is when the design parameter ϕ is close to zero that the variance bound in Theorem 2.14 tends to be loose. However, as we noted in Section 2.4, it is often beneficial to set ϕ to a value closer to one, in which case the bound is sharper.

2.7 Kernelizing the Gram–Schmidt Walk Design

One limitation of the GSW-Design presented in Section 2.3 is that the Horvitz–Thompson estimator has been shown to enjoy improved precision only when the outcomes are *linearly* related to the covariates. In many practical settings, the experimenter may believe that outcomes are better approximated by a non-linear function of the covariates. In these settings, the GSW-Design does not seem to offer a clear improvement in precision.

In this section, we show that a simple modification of the GSW-Design allows for improved precision when outcomes are related to the covariates in a non-linear way. In particular, we propose the use of *kernel methods* in the construction of the augmented covariate vectors. This kernel modification of the GSW-Design has the interpretation of lifting the covariates to a much higher dimensional vector space, where these higher dimensional vectors are balanced by the GSW-Design. Such a technique is referred to as the "kernel trick" and has been used in a variety of statistical methods including regression, classification, and unsupervised learning.

The main technical result is a generalization of the mean squared error bound of Theorem 2.14. Informally speaking, we show that by kernelizing the GSW-Design, the Horvitz–Thompson estimator enjoys improved precision when the outcomes are well-approximated by a "simple" function in the associated Reproducing Kernel Hilbert Space. Similarly, a generalization of the subgaussian tail bound for the Horvitz–Thompson estimator (Proposition 2.20) under the kernelized GSW-Design also holds, but we focus our discussion on the mean squared error in the interest of space. We remark here that the results in this section are new and do not appear in the working paper (Harshaw et al., 2021).

2.7.1 Primer on the theory of kernels and RKHS

In this section, we give a primer on the theory of kernels and Reproducing Kernel Hilbert Spaces (RKHS). A reader who is familiar with this material may wish to glance at our notation presented here and then skip to Section 2.7.2 where we present the kernelized GSW-DESIGN. Readers unfamiliar with the theory of kernels and RKHS should read this section for a high level understanding of the material. For a more complete treatment of the material, we refer readers to the textbooks (Berlinet and Thomas-Agnan, 2004; Steinwart and Christmann, 2008) and the excellent course notes of Gretton (2020), on which our primer is based.

We begin our discussion by presenting the definition of a kernel.

Definition 2.23. Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *kernel* if there exists a Hilbert space[3] $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ and a mapping $\psi : \mathcal{X} \to \mathcal{V}$ such that

$$k(\boldsymbol{x}, \boldsymbol{y}) = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y}) \rangle_{\mathcal{V}} \ .$$

Note that Definition 2.23 places virtually no restriction on the underlying set $\mathcal{X}$. For example, the set $\mathcal{X}$ may consist of discrete points, it may be a subset of $\mathbb{R}^d$, or it might be some cartesian product of the two. In our application, $\mathcal{X}$ will be the covariate space. In this way, each unit $i \in [n]$ will have an associated pre-treatment covariate $\boldsymbol{x}_i \in \mathcal{X}$.

The mapping $\psi : \mathcal{X} \to \mathcal{V}$ is referred to as the *feature map*. The feature map assigns each point in the space of covariates to a high-dimensional representation in $\mathcal{V}$. The Hilbert space $\mathcal{V}$ may be infinite dimensional, so it is infeasible to store—let alone compute—the feature mapping $\psi(\boldsymbol{x})$. However, a closed form expression of the kernel is typically available, as we will see below.

We now list a few very simple examples of kernels and we will see more complicated ones throughout the section.

- **Ex 1: Identity kernel**: Let $\mathcal{X} = \{1, 2, \dots m\}$ be a finite set. The function

  $$k(i, j) = \mathbf{1}[i = j]$$

  is the *identity kernel*. This kernel is realized by the vector space $\mathcal{V} = \mathbb{R}^n$ and the feature mapping $\psi(i) = \boldsymbol{e}_i$, where $\boldsymbol{e}_1, \boldsymbol{e}_2, \dots \boldsymbol{e}_n$ are a set of orthonormal vectors.

- **Ex 2: Linear kernel** Let $\mathcal{X} = \mathbb{R}^d$. The function

  $$k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$$

  is the *linear kernel*. This kernel is realized by the vector space $\mathcal{V} = \mathbb{R}^n$ and the identity feature mapping $\psi(\boldsymbol{x}) = \boldsymbol{x}$.

[3] A Hilbert space is an inner-product space satisfying a technical condition. Namely, that the induced metric $d_{\mathcal{V}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathcal{V}} = \sqrt{\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y} \rangle_{\mathcal{V}}}$ is complete.

- **Ex 3: A more interesting kernel** Let $\mathcal{X} = \mathbb{R}^2$. The function

$$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}(1)\boldsymbol{y}(1) + \boldsymbol{x}(2)\boldsymbol{y}(2) + \boldsymbol{x}(1)\boldsymbol{x}(2)\boldsymbol{y}(1)\boldsymbol{y}(2)$$

  is a kernel. This kernel is realized by the vector space $\mathcal{V} = \mathbb{R}^3$ and the mapping

$$\psi(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x}(1) \\ \boldsymbol{x}(2) \\ \boldsymbol{x}(1)\boldsymbol{x}(2) \end{bmatrix} .$$

For a given kernel, we remark that the feature mapping ψ and the Hilbert space $\mathcal{V}$ are not unique. At this point, it might seem like the definition of kernels is so broad that it might include any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. However, this is not so. In particular, a kernel satisfies certain structural properties, such as the inequality

$$k(\boldsymbol{x}, \boldsymbol{y})^2 \leq k(\boldsymbol{x}, \boldsymbol{x})k(\boldsymbol{y}, \boldsymbol{y}) \ ,$$

which follows from applying the Cauchy-Schwarz inequality in the Hilbert space $\mathcal{V}$.

Kernels have a whole calculus that helps us build larger kernels from smaller ones. We state a few of these rules precisely below.

Fact 2.24. *Let k_1 and k_2 be kernels on a set $\mathcal{X}$ and let α be a non-negative scalar. The following functions are also kernels on $\mathcal{X}$.*

- *(Non-negative scaling) $k = \alpha \cdot k_1$.*
- *(Addition) $k = k_1 + k_2$.*
- *(Products) $k = k_1 \cdot k_2$.*

Verifying the non-negative scaling and addition rules are straightforward, while verifying the product rule requires a bit more care. These basic ideas in Fact 2.24 can be extended to analytic functions with non-negative coefficients. Suppose that g is an analytic function that converges in the open interval $(-r, r)$, i.e.

$$g(z) = \sum_{n=0}^{\infty} a_n z^n \quad |z| < r, z \in \mathbb{R}.$$

If $\mathcal{X}$ is defined to be the open $\sqrt{r}$-ball in $\mathbb{R}^d$ and the Taylor series coefficients are nonnegative, $a_i \geq 0$, then

$$k(\boldsymbol{x}, \boldsymbol{y}) = g(\langle \boldsymbol{x}, \boldsymbol{y} \rangle) = \sum_{n=0}^{\infty} a_n \langle \boldsymbol{x}, \boldsymbol{y} \rangle^n$$

is a kernel. This is effectively a limit argument applied to the properties in Fact 2.24. The calculus in Fact 2.24 (along with analytic extensions) allows us to construct many

new kernel functions without explicitly writing down the feature mapping. We list a few here below.

- **Ex 4: Polynomial kernel**: Let $\mathcal{X} = \mathbb{R}^d$ and let m be a positive integer and $c \geq 0$. The function
$$k(\boldsymbol{x}, \boldsymbol{y}) = \Big(\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c\Big)^m$$
is the *polynomial kernel*. To see that this is indeed a kernel, observe that by the binomial theorem,
$$k(\boldsymbol{x}, \boldsymbol{y}) = \Big(\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c\Big)^m = \sum_{\ell=0}^{m} \binom{m}{\ell} \langle \boldsymbol{x}, \boldsymbol{y} \rangle^{m-\ell} c^{\ell}.$$
Because $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is a kernel and the product of kernels are kernels, we have that $\langle \boldsymbol{x}, \boldsymbol{y} \rangle^{m-\ell}$ is a kernel for each $\ell = 0, \ldots m$. The proof is completed by using the fact that the non-negative sum of kernels is a kernel.

- **Ex 5: Exponential Kernel**: Let $\mathcal{X} = \mathbb{R}^d$. The function
$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\big(\langle \boldsymbol{x}, \boldsymbol{y} \rangle\big)$$
is the *exponential kernel*. The fact that this is a kernel may be verified by applying composing the linear kernel with the exponential function, which is an analytic function with non-negative coefficients.

- **Ex 6: Gaussian Kernel**: Let $\mathcal{X} = \mathbb{R}^d$ and let $\sigma > 0$. The function
$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2}\right)$$
is the *gaussian kernel*. Verifying that this is a kernel is somewhat involved and so we omit a proof sketch.

The Gaussian kernel is a popular kernel in machine learning. The Gaussian kernel acts as a similarity function, where the similarity between a pair of points decays exponentially in their distance. The parameter σ controls this rate of decay and is referred to as the *bandwidth*.

One of the most interesting aspects of kernels is that a kernel on $\mathcal{X}$ defines a space of functions on $\mathcal{X}$, which is known as the Reproducing Kernel Hilbert Space. We state this definition formally below.

Definition 2.25. Let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be a Hilbert space of $\mathbb{R}$-valued function on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *reproducing kernel* and $\mathcal{H}$ is a *reproducing kernel hilbert space* (RKHS) if

- For all $\boldsymbol{x} \in \mathcal{X}$, $k(\cdot, \boldsymbol{x}) \in \mathcal{H}$.

- For all $\boldsymbol{x} \in \mathcal{X}$ and $f \in \mathcal{H}$, $\langle f, k(\cdot, \boldsymbol{x})\rangle_{\mathcal{H}} = f(\boldsymbol{x})$.

The first condition requires that for each element $\boldsymbol{x} \in \mathcal{X}$, the function obtained by fixing one argument of the kernel (i.e. $h_{\boldsymbol{x}}(\boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$) is in the RKHS. The second condition stipulates that evaluation of a function f in the RKHS on an element $\boldsymbol{x} \in \mathcal{X}$ is obtained by the RKHS inner product between f and $h_{\boldsymbol{x}}(\boldsymbol{y}) = k(\boldsymbol{y}, \boldsymbol{x})$. Given a kernel, k, there is a unique RKHS for which k is the reproducing kernel.

The RKHS is itself a Hilbert space and so it is equipped with the norm $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$. The RKHS norm captures complexity or irregularity of a function in the RKHS. Of course, this depends on the underlying kernel and its interpretation. To demonstrates this, we discuss two examples.

The first example is the RKHS corresponding to the linear kernel. Recall that the linear kernel is defined on $\mathcal{X} = \mathbb{R}^d$ and the kernel is the standard Euclidean inner product, $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y}\rangle$. The corresponding RKHS is

$$\mathcal{H} = \{f_{\boldsymbol{\beta}}(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x}\rangle \text{ for } \boldsymbol{\beta} \in \mathbb{R}^d\} \quad \text{with} \quad \langle f_{\boldsymbol{\beta}}, f_{\boldsymbol{\beta}'}\rangle_{\mathcal{H}} = \sum_{i=1}^{d} \boldsymbol{\beta}(i)\boldsymbol{\beta}'(i) = \langle \boldsymbol{\beta}, \boldsymbol{\beta}'\rangle \ .$$

Note that this is all linear functions on $\mathbb{R}^d$, which is isomorphic to $\mathbb{R}^d$ itself. The RKHS norm is exactly the sum of the squares of the coefficients of the linear function, i.e. $\|f_{\boldsymbol{\beta}}\|_{\mathcal{H}} = \sum_{i=1}^{n} \boldsymbol{\beta}(i)^2$. As discussed in Section 2.2.4, the sum of the squares of the coefficients of a linear function is typically used as a measure of its complexity.

The second example is the RKHS corresponding to the Gaussian kernel. For simplicity, we restrict our discussion to $d = 1$ dimension, although a similar characterization may be obtained in the multivariate setting. The Gaussian kernel admits the eigendecomposition

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right) = \sum_{\ell=1}^{\infty} \lambda_\ell \boldsymbol{e}_\ell(x)\boldsymbol{e}_\ell(x') \ ,$$

where $\lambda_\ell = b^\ell$ for b, a constant which is increasing with the bandwidth σ and $\boldsymbol{e}_\ell(x) = \exp(-\gamma x^2)H_\ell(c \cdot x)$, where H_ℓ is the ℓth-order Hermite polynomial and c and γ are constants depending on σ (see, e.g. Section 4.3 in Rasmussen and Williams, 2006). As ℓ increases, the basis functions $\boldsymbol{e}_\ell(x)$ increase in complexity. Let L_2 be the set of real-valued square integrable functions on $\mathbb{R}$ with respect to the Gaussian measure. Functions $f, g \in L_2$ admit an expansion in an orthonormal system $\{\boldsymbol{e}_\ell\}_{\ell=1}^{\infty}$,

$$f(x) = \sum_{\ell=1}^{\infty} \hat{f}_\ell \boldsymbol{e}_\ell(x) \quad g(x) = \sum_{\ell=1}^{\infty} \hat{g}_\ell \boldsymbol{e}_\ell(x)$$

and the standard inner product on L_2 is defined as

$$\langle f, g \rangle_{L_2} = \Big\langle \sum_{\ell=1}^{\infty} \hat{f}_\ell \boldsymbol{e}_\ell(x), \sum_{\ell=1}^{\infty} \hat{g}_\ell \boldsymbol{e}_\ell(x) \Big\rangle_{L_2} = \sum_{\ell=1}^{\infty} \hat{f}_\ell \hat{g}_\ell \ .$$

We are now ready to describe the RKHS norm with respect to the Gaussian kernel. The inner product is similar to the standard L_2 inner product, except that it features a penalty on the roughness / irregularity on higher order terms in the basis expansion:

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell \hat{g}_\ell}{\lambda_\ell} \quad \text{inducing the norm} \quad \|f\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{\infty} \frac{\hat{f}_\ell^2}{\lambda_\ell} \ .$$

The RKHS is the subset of functions L_2 for which the norm $\|f\|_{\mathcal{H}}^2$ converges. Note that the norm converges only when the higher order terms in the basis expansion decay at a rate faster than $\lambda_\ell = b^\ell$. In this sense, a function of low RKHS norm is simple, as it does not involve a significant amount of higher order terms.

A variety of interesting kernels have been developed in the machine learning community for specific applications. Several kernels have been proposed for data which is represented by a graph (Kondor and Lafferty, 2002; Smola and Kondor, 2003; Vishwanathan et al., 2010). The kernels have interpretations of capturing random walks or diffusion processes on the graph and the RKHS norm typically measures the irregularity of the function with respect to the edges in the graph. Another interesting type of kernel is the Neural Tangent Kernel, which is obtained as the infinite-width limit of a neural network trained on a squared loss using gradient descent (Jacot et al., 2018). These kernels were proposed to provide insight to training dynamics and generalization properties of the neural network. It is possible that experimenters may benefit from these recent advances in kernel methods, but such a connection is beyond the scope of this chapter.

2.7.2 The kernelized GSW-Design

We assume that the experimenter has already collected pre-treatment covariates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n \in \mathcal{X}$ for each unit and that a kernel k has been selected. We remark that the covariates are not limited in any structural way. In particular, the covariates do not need to be vectors in $\mathbb{R}^d$. Categorical and numerical data may be used together, so long as an appropriate kernel is defined. How experimenters should choose the kernel is beyond the scope of this work, although prior substantive knowledge such as a pilot study or a generative model of outcomes should be used in determining this kernel.

Whereas the GSW-Design attempted to balance raw covariate vectors $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n \in \mathbb{R}^d$ in a linear sense, the kernelized GSW-Design attempts to balance the feature vectors $\psi(\boldsymbol{x}_1), \psi(\boldsymbol{x}_2), \ldots \psi(\boldsymbol{x}_n) \in \mathcal{V}$ in a linear sense. Because the feature mapping may be non-linear, balancing these higher dimensional feature vectors in a linear way

results in balancing the original raw covariate vectors in non-linear ways. The only modification in the kernelized GSW-Design is the construction of the augmented covariates. In particular, we would like to construct the *kernelized* augmented covariates $\boldsymbol{b}_1, \boldsymbol{b}_2, \dots \boldsymbol{b}_n \in \mathbb{R}^n \times \mathcal{V}$ as

$$\boldsymbol{b}_i = \begin{bmatrix} \sqrt{\phi}\boldsymbol{e}_i \\ \frac{\sqrt{1-\phi}}{\xi}\psi(\boldsymbol{x}_i) \end{bmatrix} \in \mathbb{R}^n \times \mathcal{V} \ ,$$

where $\xi = \max_{i\in[n]}\|\psi(\boldsymbol{x}_i)\|$ is the largest norm of the feature vectors. These augmented covariates would be ideal as they directly balance the feature vectors in a linear manner. Unfortunately, the feature embedding $\psi(\boldsymbol{x}_i)$ is very high (possibly infinite) dimensional and so we cannot store or compute them. Here, we use the so-called "kernel trick" which is the following observation: the GSW-Design only requires knowing the inner products between all augmented covariate vectors, which may be computed as

$$\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = \Big(\frac{1-\phi}{\xi}\Big)\Big\langle \psi(\boldsymbol{x}_i), \psi(\boldsymbol{x}_j)\Big\rangle_{\mathcal{V}} = \Big(\frac{1-\phi}{\xi}\Big)k(\boldsymbol{x}_i, \boldsymbol{x}_j) \ .$$

With this kernel trick in mind, we now describe a more mechanical way to sample an assignment from the kernelized GSW-Design. Let $\boldsymbol{K}$ be the symmetric n-by-n symmetric kernel matrix whose (i,j)th entry is the kernel evaluated on the ith and jth covariates, i.e. $\boldsymbol{K}(i,j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Let $\xi^2 = \max_{i\in[n]} k(\boldsymbol{x}_i, \boldsymbol{x}_i)$ be the maximum kernel evaluation over the covariates. Next, construct a matrix factorization $\boldsymbol{K} = \boldsymbol{M}^\intercal\boldsymbol{M}$, where M is n-by-n with columns $\boldsymbol{m}_1, \boldsymbol{m}_2, \dots \boldsymbol{m}_n \in \mathbb{R}^n$. Given the design parameter $\phi \in [0,1]$, we define the *kernelized* augmented covariate vectors as

$$\boldsymbol{b}_i = \begin{bmatrix} \sqrt{\phi}\boldsymbol{e}_i \\ \frac{\sqrt{1-\phi}}{\xi}\boldsymbol{m}_i \end{bmatrix} \in \mathbb{R}^{2n} \ .$$

By construction, the vectors $\boldsymbol{m}_1, \boldsymbol{m}_2, \dots \boldsymbol{m}_n$ satisfy the property that $\langle \boldsymbol{m}_i, \boldsymbol{m}_j \rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ so that the inner product between augmented covariate vectors is $\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = (1-\phi)/\xi \cdot k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, as desired. As assignment is drawn from the kernelized GSW-Design by running the GSW-Design on these augmented covariate vectors.

Using the computational techniques discussed in Section 2.3.2 (and described in detail in Section A.2), a sample may be drawn using $\mathcal{O}(n^3)$ arithmetic operations, as the vectors $\boldsymbol{m}_1, \boldsymbol{m}_2, \dots \boldsymbol{m}_n$ are generally n-dimensional. Moreover, an additional one-time cost of $\mathcal{O}(n^2)$ kernel evaluations is required to construct the kernel matrix $\boldsymbol{K}$. It is possible that the kernel matrix may be well-approximated using fewer kernel evaluations either by Nyström sampling (Drineas and Mahoney, 2005; Gittens and Mahoney, 2013) or random projection (Yang et al., 2017) methods. However, the computational speed up will generally be inconsequential to experimenters who may view the construction of the kernel matrix as a one-time pre-processing cost.

2.7.3 Analysis of the mean squared error

We now present the main result of this section, which is a bound on the mean squared error of the Horvitz–Thompson estimator under the kernelized GSW-Design. The proof of this theorem is similar to that of Theorem 2.14, except that it makes use of the representer theorem (Schölkopf et al., 2001) to obtain the kernel ridge regression loss. The proof appears in Appendix A.3.5.

Theorem 2.26. *Let $\mathcal{X}$ be the space of covariates, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel on the covariates, and let $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ be the associated RKHS. The mean squared error of the Horvitz–Thompson estimator under the kernelized* GSW-Design *is at most the minimum of the loss function of an implicit kernel ridge regression of the sum of the potential outcomes on the covariates:*

$$\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] \leq \frac{1}{n} \cdot \min_{f \in \mathcal{H}} \left[\frac{1}{\phi} \cdot \frac{1}{n} \sum_{i=1}^{n} \big((a_i + b_i) - f(\boldsymbol{x}_i)\big)^2 + \frac{\xi^2}{(1-\phi)n} \big\| f \big\|_{\mathcal{H}}^2 \right].$$

In many ways, Theorem 2.26 resembles Theorem 2.14. The design parameter $\phi \in [0, 1]$ trades off the emphasis between two terms in the objective: one which captures the fit of the regression and the second which captures its complexity. The first term measures how well the outcomes are approximated by the function of the covariates. The second term is the RKHS norm of the approximator, which (for most kernels) is considered a measure of the irregularity of complexity of the function. In this way, Theorem 2.26 demonstrates that the Horvitz–Thompson estimator achieves high precision under the kernelized GSW-Design when the outcomes are well-explained by a simple function in the RKHS defined by the kernel. This extends the "regression-by-design" result to non-linear regressions.

2.8 Conclusion and Open Problems

There are several open questions suggested by this work. Answering any of the following methodological questions would shed light on the nature of the balance-robustness trade-off in ways we have not yet explored here.

- **Instance-Optimal Subguassian Bound**: *Show the subgaussian bound on the Horvitz–Thompson estimator under the* GSW-Design *(Theorem 2.20) holds when L/n is replaced by the variance of the estimator,* $\mathrm{Var}(\widehat{\tau})$. We have demonstrated a subgaussian tail bound on the Horvitz–Thompson estimator under the GSW-Design with L/n in the denominator, which is an upper bound on the variance; replacing this upper bound with the true variance would yield tighter tail bounds, especially when ϕ is further away from 1. Currently, the subgaussian bound on the Horvitz–Thompson estimator follows from the subgaussian bound of the discrepancy vector returned by the Gram–Schmidt Walk algorithm, which is $\sigma^2 \leq \max_{i \in [n]} \|\boldsymbol{b}_i\|^2$ (Theorem 2.19). In general, this subgaussian

constant σ^2 is tight (e.g. on orthogonal input vectors) and so improving the subgaussian analysis of the discrepancy vector would require that σ^2 depends on all the input vectors in a finer way. Any improved analysis would use very different techniques than those presented here.

- **Asymptotic Normality**: *Prove that under certain regularity conditions on the potential outcomes, the Horvitz–Thompson estimator is asymptotically normal under the* GSW-Design. This would motivate the use of smaller confidence intervals in large samples based on a normal approximation, which would be asymptotically valid and are typically smaller than intervals obtained from the subgaussian bound. The main challenge here is that the GSW-Design introduces minor amounts of dependence between all assignments. This precludes establishing central limit theorems using modern techniques, such as Stein's method via dependency graphs, as the degree of the dependency graph in this setting would grow too rapidly with n.

- **Online Covariate Balancing**: *Construct an online algorithm for Problem 2.5 where C is a constant, independent of n and d—or show that no such algorithm exists.* In this chapter, we considered the setting where all covariate vectors are known prior to assignment; however, it is common in many real-world settings that units need to be assigned in a sequential manner as they arrive to the study. In these settings, one cannot directly apply the GSW-Design. In the *online* setting, the experimenter observes covariate vectors in a sequence $\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_T$. At each iteration t, the experimenter must randomly assign treatment $z_t \in \{\pm 1\}$ after observing all previous covariate vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_t$ and before observing the next covariate vector $\boldsymbol{x}_{t+1}$. Constructing a design which navigates the balance-robustness trade-off in the online setting would be valuable to experimenters as well as interesting in the growing literature on online discrepancy algorithms.

Chapter 3

Optimized Variance Estimation under Interference and Complex Experimental Designs [1]

3.1 Introduction

The design-based approach to causal inference considers random assignment of treatments as the only source of randomness. In this framework, the variance of treatment effect estimators depends on products of potential outcomes. But such products cannot be estimated well, even under well-behaved designs, because some pairs of potential outcomes are never observed at the same time. Without additional assumptions, unbiased and consistent variance estimation is not possible in the design-based setting.

To demonstrate this, we present a very simple example. Consider an experiment with $n = 1$ unit whose outcomes under treatment ($z_1 = 1$) and control ($z_1 = -1$) are a_1 and b_1, respectively. The treatment effect is the contrast of these two outcomes: $\tau = a_1 - b_1$. Under the Bernoulli design where $\Pr(z_1 = 1) = 1/2$, the Horvitz–Thompson estimator is simply $\widehat{\tau} = 2z_1y_1$, where y_1 is the observed outcome. The Horvitz–Thompson estimator is unbiased, $\mathbb{E}[\widehat{\tau}] = \tau$ and its variance is

$$\mathrm{Var}(\widehat{\tau}) = a_1^2 + b_1^2 + 2a_1b_1 \ .$$

Although the terms a_1^2 and b_1^2 may be estimated without bias, the product a_1b_1 is never observed. In fact, no information from one run of the experiment can inform us about even the sign of this term a_1b_1. This problem remains *even as the sample size grows*: indeed, there exists neither an unbiased nor a consistent estimator of the variance (Imbens and Rubin, 2015).

[1] Based on the working paper: Christopher Harshaw, Joel Middleton, and Fredrik Sävje (2021) "Optimized Variance Estimation under Interference and Complex Experimental Designs". Forthcoming.

The variance estimation problem is more challenging in the presence of interference and under complex designs. When the experiment exhibits interference (defined formally in Section 3.2.1), some pairs of potential outcomes between units are also necessarily unobserved. For example, interference arises in social network experiments, where a unit can be directly treated (treatment is received), indirectly treated (a neighbor receives treatment) or untreated (the unit nor its neighbors receive treatment). If a unit is treated, then all of its neighbors must receive either direct or indirect treatment. In this way, two neighboring units cannot receive direct treatment and control, which exacerbates the problem of variance estimation. Complex designs, such as matched-pair designs, cluster randomization, and various block designs, restrict the treatment assignment so that two units are never assigned a specific configuration of treatments, which also further exacerbates the variance estimation problem.

The goal of this chapter is to address the variance estimation problem under complex experimental designs by constructing improved variance estimators. Unlike previous works (reviewed in Section 3.1.2) which construct variance estimators in specific settings, we consider arbitrary interference, arbitrary designs, and a large class of linear point estimators that includes all commonly used treatment effect estimators. Our work is centered on obtaining *variance bounds*, which are upper bounds on the variance which themselves admit unbiased and consistent estimators. The main contributions are summarized as follows:

- In Section 3.2, we describe and characterize the variance estimation problem under arbitrary interference for arbitrary designs for the full class of linear estimators. In particular, we characterize those variance bounds which are admissible for the problem of variance estimation.

- In Section 3.3, we propose an optimization based framework, termed OPT-VB, for selecting an admissible variance bound. We describe methods which allow an experimenter to select a variance bound based on their level of risk aversion and prior substantive knowledge about the unknown potential outcomes. An algorithm for testing admissibility of a variance bound is described in Section 3.4.

- In Section 3.5, we discuss how a variance estimator may be obtained from a variance bound. In particular, we describe a Horvitz–Thompson estimator of the variance bound and give conditions under which it is consistent. We argue that this consistency condition may inform the choice of variance bound and the setting of OPT-VB.

3.1.1 Variance bounds: an illustration

In this section, we provide a stylized experimental setting with interference to illustrate the concept of variance bounds. A formal discussion of the general experimental setting is deferred to Section 3.2.

Social scientists are interested in how people's behavior is affected by information. A potentially important aspect is how the information is transmitted. Information from a credible, first-hand source might be more effective than second-hand information. Consider a study that investigates this in the context of political campaigning. Campaigns often reach out to potential voters in an effort to persuade them to vote for a particular candidate. People targeted by a campaign might in turn spread the message to people who were not directly targeted by the campaign. Our interest in this illustration is to estimate the difference in voting behavior when being directly targeted by the campaign and when being only indirectly targeted by knowing someone who is directly targeted.

Consider a sample of two potential voters, who are units in the experiment indexed by $i \in \{1, 2\}$. The two voters might be part of a bigger social network, but we restrict our attention to this small sample in this illustration for simplicity. There are two experimental conditions: either voter 1 is directly targeted by the campaign and person 2 is indirectly targeted (through conversations with person 1) or vice versa, i.e. person 2 is directly targeted and person 1 is indirectly targeted. For each person $i \in \{1, 2\}$ let a_i denote the voting behavior of person i under direct targeting, and let b_i denote voting behavior of the same person when indirectly targeted. The causal quantity of interest is average contrast in voting behavior between direct and indirect campaign targeting, $\tau = 1/2 \cdot (a_1 - b_1) + 1/2 \cdot (a_2 - b_2)$. The experimental design stipulates that each experimental condition occurs with probability 1/2. In this case, the Horvitz–Thompson estimator has the distribution

$$\widehat{\tau} = \begin{cases} a_1 - b_2 & \text{with probability } 1/2, \\ a_2 - b_1 & \text{with probability } 1/2. \end{cases}$$

Thus, the variance of the estimator is

$$\mathrm{Var}(\widehat{\tau}) = \frac{1}{4}(a_1^2 + a_2^2 + b_1^2 + b_2^2) + \frac{1}{2}(a_1 b_1 + a_2 b_2 - a_1 a_2 - b_1 b_2 - a_1 b_2 - a_2 b_1).$$

The heart of the problem is that some terms in the variance expression are never observed. We never observe $a_1 b_1$ or $a_2 b_2$, because a person is never directly and indirectly targeted by the campaign simultaneously. Similarly, we never observe $a_1 a_2$ or $b_1 b_2$, because Person 1 is directly targeted when Person 2 is indirectly targeted, and vice versa. The unobserved terms prevent us from constructing an unbiased estimator of the variance unless we make assumptions about the outcomes. The assumptions we would need to make are strong, and they are often untenable in practice.

An alternative route is to construct a bound for the variance, which allows us to construct a conservative estimator. A simple bound uses the fact that

$$\mathrm{Var}(\widehat{\tau}) = \mathbb{E}[\widehat{\tau}^2] - \mathbb{E}[\widehat{\tau}]^2 \leq \mathbb{E}[\widehat{\tau}^2].$$

Hence, the variance is upper bounded by

$$\mathrm{Var}(\widehat{\tau}) \leq \mathbb{E}\big[\widehat{\tau}^2\big] = \frac{1}{2}(a_1^2 + a_2^2 + b_1^2 + b_2^2) - a_1 b_2 - a_2 b_1.$$

Note that this bound holds for any values of outcomes, so no assumptions are required here. Furthermore, the bound is estimable because all terms are observed with some positive probability. Although, no estimator will be able to precisely estimate the variance in this example due to the small sample size.

The bound just derived is just one of many possible bounds. A somewhat more intricate bound uses the fact that

$$(x - y)^2 = x^2 - 2xy + y^2 \geq 0,$$

which means that $(x^2 + y^2)/2$ is an upper bound for the product xy for any real-valued x and y. This is known as the Arithmetic-Geometric (AM-GM) inequality or Young's inequality. Applying this inequality to the problematic terms of the variance, we arrive at

$$\mathrm{Var}(\widehat{\tau}) \leq \frac{3}{4}(a_1^2 + a_2^2 + b_1^2 + b_2^2) - \frac{1}{2}(a_1 b_2 + a_2 b_1).$$

These terms are the same terms as above but with different coefficients. Hence, they are all observed with some positive probability, and the second bound is also estimable.

Both of these bounds are valid and estimable, so either can be used to construct a variance estimator that is conservative in expectation. Indeed, there are infinitely many variance bounds here, with infinitely many corresponding conservative variance estimators. The question we ask in this chapter is which of these bounds we should use. We want a variance estimator that is guaranteed to be conservative, which would be achieved by any one of the bounds, but we want to avoid excessive conservativeness if possible.

The choice is simple when choosing between the two bounds in this illustration, because the first bound is always smaller than the second; that is, the second bound is inadmissible. More generally, however, the set of admissible bounds will be infinitely large, and there is no universal ordering among them. We suggest that experimenters take advantage of background information about the potential outcomes when selecting a bound to use to construct a variance estimator. In what follows, we describe how this can be achieved while ensuring that the resulting estimator is conservative and estimable no matter if the supplied background information is correct.

3.1.2 Related works

To the best of our knowledge, Neyman (1923) was the first to recognize that variances of treatment effect estimators are not directly estimable. He showed that the variance of the difference-in-means estimator under the complete randomization de-

sign depends on the covariance of unit-level potential outcomes, which cannot be estimated from the data. Neyman applied the Cauchy–Schwarz inequality followed by the AM–GM inequality to arrive at an upper bound of the variance that could be estimated.

Neyman's approach has been improved and extended in several directions. An important line of work considers variance estimation under other experimental designs than complete randomization. Early examples include Kempthorne (1955) and Wilk (1955), who studied variance estimation under various blocked designs. However, these investigations generally impose structural assumptions on the potential outcomes, such as constant treatment effects, which limits its applicability. The more recent literature has derived Neyman-type variance estimators for a large class of designs without such assumptions (see, e.g., Gadbury, 2001; Abadie and Imbens, 2008; Imai, 2008; Higgins et al., 2015; Fogarty, 2018; Pashley and Miratrix, 2019).

A related strand of the literature has derived Neyman-type variance estimators for other point estimators than the difference-in-means estimator. For example, Samii and Aronow (2012) and Aronow and Middleton (2013) investigate variance estimators for the ordinary least square regression estimator and the Horvitz–Thompson estimator, respectively. To the best of our knowledge, variance estimation for general estimators under arbitrary designs has not previous been studied.

Another strand of the literature aims to sharpen Neyman's bound. Robins (1988) focuses on binary outcomes and derives a variance estimator that extracts all information about the joint distribution of the potential outcomes contained in the marginal distributions. Aronow et al. (2014) use Fréchet–Hoeffding-type bounds to generalize the estimator by Robins (1988) to arbitrary outcome variables. Nutz and Wang (2020) provide further improvements under the assumption that all unit-level treatment effects are nonnegative. Menzel and Imbens (2021) provide higher-order refinements to these bounds using a bootstrap approach. While these bounds can be useful when experimenters use the difference-in-means estimator under complete randomization, it is unclear whether and how the results generalize to more complex estimators and designs.

The strand of the literature closest to our contribution considers variance estimation for arbitrary designs. To the best of our knowledge, the only previous result here is due to Aronow and Samii (2013, 2017). They describe a method to construct a bound for the variance of the Horvitz–Thompson estimator when some pair-wise assignment probabilities are zero. However, as we noted in the previous section, there are generally an infinite number of admissible variance bound. We here explore whether there exist better bounds. Indeed, there does—the Aronow–Samii bound is often inadmissible.

3.2 Linear Point Estimators and their Variance

3.2.1 Preliminaries

The focus of our study is an experiment consisting of n units indexed by $[n] = \{1, \ldots, n\}$. The experimenter randomly assigns treatment $z_i \in \{0, 1\}$ to each unit $i \in [n]$ and we collect these assignments into the random vector $\boldsymbol{z} = (z_1, \ldots z_n) \in \{0, 1\}^n$. The distribution of $\boldsymbol{z}$ is the *design* of the experiment, which is selected by, and thus known to, the experimenter. Each unit is associated with a vector of d deterministic covariates $\boldsymbol{x}_i \in \mathbb{R}^d$, which are also known the experimenter. The matrix produced by stacking $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as rows is denoted $\boldsymbol{X}$.

Each unit $i \in [n]$ has an associated *potential outcome* function $y_i : \{0, 1\}^n \to \mathbb{R}$ that specifies the response of unit i under all possible treatment assignments. This allows the response of unit i to depend not only on its own treatment, but potentially also on the treatment of other units. In other words, we allow for interference. The potential outcomes themselves are deterministic and the only randomness in the observed outcomes arises from randomness in the treatment assignment. In particular, the observed outcome of unit i is $y_i(\boldsymbol{z})$, which is random because it depends on the treatment assignments. In an abuse of notation, we sometimes use y_i to refer to both the (deterministic) potential outcome function as well as the (random) observed outcome, formally written as $y_i(\boldsymbol{z})$. The random vector of observed outcomes is denoted $\boldsymbol{y} = (y_1, y_2, \ldots y_n)$.

To model interference in the potential outcomes framework, we will use exposure mappings as described by Aronow and Samii (2017). In this framework, the exposures capture the relevant information required to specify the outcomes. More formally, each unit $i \in [n]$ has an *exposure mapping* $d_i : \{0, 1\}^n \to \Delta$ that maps the treatment assignment to a finite set of exposures Δ. The assignment vectors mapping to the same exposure for some unit are considered causally similar or equivalent with respect to that unit. The number of exposures $|\Delta|$ is typically considered to be small with respect to the number of units. For example, in a social network experiment we may define $\Delta = \{e_{\text{direct}}, e_{\text{indirect}}, e_{\text{control}}\}$ which corresponds to three possible exposure levels. The exposure mapping for unit $i \in [n]$ may be given by

$$d_i(\boldsymbol{z}) = \begin{cases} e_{\text{direct}} & \text{if } z_i = 1 \\ e_{\text{indirect}} & \text{if } z_i = 0 \text{ and } z_j = 1 \text{ for some neighbors } j \text{ of } i \\ e_{\text{control}} & \text{if } z_i = 0 \text{ and } z_j = 0 \text{ for all neighbors } j \text{ of } i \end{cases}.$$

Researchers using exposure mappings to model interference often assume the mappings are correctly specified, and we will assume the same in this chapter. The assumption states that the outcomes are completely determined by the exposure received by each unit, so that

$$y_i(\boldsymbol{z}) = y_i(\boldsymbol{z}') \quad \text{for all} \quad \boldsymbol{z}, \boldsymbol{z}' \in \{0, 1\}^n \quad \text{such that} \quad d_i(\boldsymbol{z}) = d_i(\boldsymbol{z}').$$

In other words, if two assignment vectors produce the same exposure for unit i, then the outcome of unit i will be the same for these two assignment vectors. This means that all relevant causal quantities can be understood and defined in terms of the exposures received by the units. For each unit $i \in [n]$, we define the random variable $D_i = d_i(\boldsymbol{z})$ as the exposure which that unit receives.

The causal quantities of interest when using exposure mappings are generally the average contrast between two exposures for all units. Given two exposures $a, b \in \Delta$, we consider the estimand

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

where we have overloaded notation by writing $y_i(d)$ to denote the outcome of unit i under exposure $d \in \Delta$. This class of estimand includes many commonly studied causal quantities, including total, direct and indirect treatment effects. In the remainder of the chapter, we write $\tau(a, b)$ simply as τ for notational convenience. Although we focus on estimands that are unweighted averages of contrasts of potential outcomes in this chapter, the results easily generalize to causal estimands which are arbitrary linear functions of the potential outcomes.

To recover the conventional no-interference setting, one uses exposure mappings that set each unit's exposure to its own treatment, $d_i(\boldsymbol{z}) = z_i$, which gives a binary set of exposures, $\Delta = \{0, 1\}$. The conventional average treatment effect is then produced by setting $a = 1$ and $b = 0$ in the contrast of the estimand. Therefore, the investigation in this chapter is applicable to the no-interference setting without modification.

3.2.2 Linear treatment effect estimators

We consider the class of *linear estimators*. Estimators in this class can be written as random linear functions of the observed outcomes:

$$\widehat{\tau} = \sum_{i=1}^{n} w_i y_i, \tag{3.1}$$

where the coefficients w_i may depend arbitrarily on treatment assignment $\boldsymbol{z}$ and covariates $\boldsymbol{X}$. Thus, the coefficients w_i can be random, but they cannot depend on the outcomes. While the estimators in this class are linear functions, they do not require or implicitly impose any linearity assumption on the conditional expectation functions of the outcomes given the covariates; that is, we can use a linear estimator without assuming a linear regression model.

This class includes most estimators commonly used by experimenters to estimate exposure effects. As the following examples demonstrate, the class includes all point estimators discussed by Aronow and Samii (2017), and many others.

1. The Horvitz–Thompson estimator (Horvitz and Thompson, 1952) uses inverse probability weighting to account for non-uniform assignment probabilities. We can write this estimator in the form of Eq. (3.1) by using the coefficients

$$w_i = \frac{\mathbf{1}[D_i = a]}{n \Pr(D_i = a)} - \frac{\mathbf{1}[D_i = b]}{n \Pr(D_i = b)}.$$

2. The difference-in-means estimator (Imbens and Rubin, 2015) contrasts the sample means between the two groups which received the exposures of interest. We can write this estimator in the linear form by using the coefficients

$$w_i = \frac{\mathbf{1}[D_i = a]}{\sum_{j=1}^{n} \mathbf{1}[D_j = a]} - \frac{\mathbf{1}[D_i = b]}{\sum_{j=1}^{n} \mathbf{1}[D_j = b]}.$$

3. The Hájek estimator (Hájek, 1971) is a generalization of the difference-in-means estimator that accommodates non-uniform assignment probabilities. We can write this estimator in the linear form by using the coefficients

$$w_i = \left(\frac{\mathbf{1}[D_i = a]}{\Pr(D_i = a)} \Big/ \sum_{j=1}^{n} \frac{\mathbf{1}[D_j = a]}{\Pr(D_j = a)} \right) - \left(\frac{\mathbf{1}[D_i = b]}{\Pr(D_i = b)} \Big/ \sum_{j=1}^{n} \frac{\mathbf{1}[D_j = b]}{\Pr(D_j = b)} \right).$$

4. The conventional OLS regression estimator of the average treatment effect (see, e.g., Duflo et al., 2007) is obtained by using the coefficients

$$w_i = \boldsymbol{e}_2^\intercal \big(\boldsymbol{Q}^\intercal \boldsymbol{Q}\big)^{-1} \boldsymbol{Q}^\intercal \boldsymbol{e}_i,$$

   where $\boldsymbol{e}_i$ is the ith standard basis vector of appropriate dimension, and $\boldsymbol{Q} = \big[\mathbf{1}, \boldsymbol{z}, \boldsymbol{X}\big]$. This estimator has been shown to perform poorly in some situations. Lin (2013) describes a modified OLS regression estimator that addresses the issue. This estimator has the same form but with the matrix $\boldsymbol{Q} = \big[\mathbf{1}, \boldsymbol{z}, \boldsymbol{X}_{\text{DM}}, \boldsymbol{X}_{\text{INT}}\big]$, where the matrix $\boldsymbol{X}_{\text{DM}} = \boldsymbol{X} - n^{-1}\mathbf{1}\mathbf{1}^\intercal \boldsymbol{X}$ is the demeaned covariate matrix, and $\boldsymbol{X}_{\text{INT}} = \boldsymbol{z}\mathbf{1}^\intercal \circ \boldsymbol{X}_{\text{DM}}$ is the Hadamard product between the treatment vector and the demeaned covariate matrix.

5. The Generalized Regression Estimator (Cassel et al., 1976), which sometimes is called the Augmented Inverse Propensity Weighted Estimator, allows for both covariate adjustment and non-uniform assignment probabilities. The estimator is written as

$$\widehat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{x}_i^\intercal \big(\widehat{\boldsymbol{\beta}}_a - \widehat{\boldsymbol{\beta}}_b\big) + \frac{\mathbf{1}[D_i = a]\big(y_i - \boldsymbol{x}_i^\intercal \widehat{\boldsymbol{\beta}}_a\big)}{\Pr(D_i = a)} - \frac{\mathbf{1}[D_j = b]\big(y_i - \boldsymbol{x}_i^\intercal \widehat{\boldsymbol{\beta}}_b\big)}{\Pr(D_j = b)} \right],$$

   where the linear functions $\widehat{\boldsymbol{\beta}}_d$ are chosen to minimize $\sum_{i=1}^{n} \mathbf{1}[D_i = d](y_i - \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_d)^2$.

We can write $\widehat{\boldsymbol{\beta}}_d$ in closed form as $\left(\boldsymbol{X}_d^\intercal \boldsymbol{X}_d\right)^{-1}\boldsymbol{X}_d^\intercal \boldsymbol{y}$, where the ith row of $\boldsymbol{X}_d$ is equal to $\boldsymbol{x}_i$ if $D_i = d$ and otherwise equal to zero. This means that we can write the estimator in the linear form by using the coefficients

$$w_i = \frac{\mathbf{1}[D_i = a]}{n \Pr(D_i = a)} - \frac{\mathbf{1}[D_i = b]}{n \Pr(D_i = b)} + \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{Q}_j^\intercal \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ is the ith standard basis vector of dimension n, and

$$\boldsymbol{Q}_j^\intercal = \left(1 - \frac{\mathbf{1}[D_j = a]}{\Pr(D_j = a)}\right)\boldsymbol{x}_j^\intercal \left(\boldsymbol{X}_a^\intercal \boldsymbol{X}_a\right)^{-1}\boldsymbol{X}_a^\intercal - \left(1 - \frac{\mathbf{1}[D_j = b]}{\Pr(D_j = b)}\right)\boldsymbol{x}_j^\intercal \left(\boldsymbol{X}_b^\intercal \boldsymbol{X}_b\right)^{-1}\boldsymbol{X}_b^\intercal.$$

Of course, the class of linear estimators contains many more members than these examples. In the remainder of the chapter, we will focus on a generic linear estimator rather than any particular estimator among these examples. As they all belong to the studied class, the results apply to all of them.

3.2.3 The variance of linear estimators

For expositional reasons, we restrict our attention in the main part of the chapter to linear estimators that depend only on observed potential outcomes corresponding to the two exposures of interest, a and b. All estimators listed in the previous section are of this type, but the class of linear estimators is larger than that. The extension to the full class is straightforward, but heavy on notation, so we relegate this discussion to Appendix B.1.

Linear estimators of this type can be written as

$$\widehat{\tau} = \sum_{i=1}^{n} \mathbf{1}[D_i = a] w_i y_i(a) + \sum_{i=1}^{n} \mathbf{1}[D_i = b] w_i y_i(b).$$

To make this expression more manageable, we extend the index set to $P = \{1, \dots 2n\}$, and for $k \in P$, we define two variables:

$$v_k = \begin{cases} \mathbf{1}[D_k = a] w_k & \text{if } k \leq n, \\ \mathbf{1}[D_{k-n} = b] w_{k-n} & \text{if } k > n, \end{cases} \qquad \text{and} \qquad \theta_k = \begin{cases} y_k(a) & \text{if } k \leq n, \\ y_{k-n}(b) & \text{if } k > n. \end{cases}$$

Collecting these variables in vectors, $\boldsymbol{v} = (v_1, \dots v_{2n})$ and $\boldsymbol{\theta} = (\theta_1, \dots \theta_{2n})$, we can write the estimator as

$$\widehat{\tau} = \sum_{k=1}^{2n} v_k \theta_k = \boldsymbol{v}^\intercal \boldsymbol{\theta}.$$

The advantage of writing the estimator in this form is that we have isolated the potential outcomes in $\boldsymbol{\theta}$, which is nonrandom, so all randomness is concentrated in

the coefficient vector $\boldsymbol{v}$. This makes the derivation of the variance of the estimator straightforward.

Lemma 3.1. *The variance of a linear estimator* $\widehat{\tau} = \boldsymbol{v}^\intercal \boldsymbol{\theta}$ *is*

$$\operatorname{Var}(\widehat{\tau}) = \boldsymbol{\theta}^\intercal \boldsymbol{A} \boldsymbol{\theta},$$

where $\boldsymbol{A} = \operatorname{Cov}(\boldsymbol{v})$ *is the covariance matrix of the coefficient vector* $\boldsymbol{v}$.

Proof. Because $\boldsymbol{\theta}$ is nonrandom, $\operatorname{Var}(\widehat{\tau}) = \operatorname{Var}\big(\boldsymbol{v}^\intercal \boldsymbol{\theta}\big) = \boldsymbol{\theta}^\intercal \operatorname{Cov}(\boldsymbol{v}) \boldsymbol{\theta}$. □

The lemma is useful because $\boldsymbol{A} = \operatorname{Cov}(\boldsymbol{v})$ does not depend on the potential outcomes, so it is known, at least in principle. Furthermore, because $\boldsymbol{A}$ is a covariance matrix, it is positive semidefinite. This means that the variance is a known positive semidefinite quadratic form of the potential outcome vector, which in turn makes it conducive to analysis.

A possible complication here is that the covariance matrix may be hard to compute for some estimators and designs. If that turns out to be the case, experimenters can then use a Monte Carlo approach to estimate the matrix (Fattorini, 2006). This generally does not cause troubles because experimenters can run the simulation until the matrix is known to desired precision. But to ease the exposition, we will proceed under the assumption that $\boldsymbol{A}$ is known.

3.2.4 The variance of linear estimators is not estimable

The task of estimating the variance of a linear estimator has now been reduced to the task of estimating the corresponding (known) quadratic form in the (unknown) potential outcome vector $\boldsymbol{\theta}$. The central concern here is that some quadratic forms cannot be estimated well. In particular, some pairs of potential outcomes may never, or only very rarely, be observed at the same time, and this can make the task difficult or impossible.

The experimental design determines which exposures are simultaneously realizable, which in turn determines which outcomes are simultaneously observable. Consider the random subset $S \subset P$ given by

$$S = \{i : D_i = a\} \cup \{i + n : D_i = b\},$$

collecting the indices of the observed elements of $\boldsymbol{\theta}$. If $\Pr(k, \ell \in S) = 0$ for some pair $\ell, k \in P$, then the corresponding product $\theta_k \theta_\ell$ is never observed. These pairs will be central to our discussion, so we let

$$\Omega = \big\{(\ell, k) : \Pr(\ell, k \in S) = 0\big\}$$

be the set of all pairs of the relevant contrasted potential outcomes that are never simultaneously observed. As formalized in the following definition and proposition,

estimable quadratic forms are those that are compatible with this pattern of observability.

Definition 3.2. An arbitrary quadratic form $\boldsymbol{\theta}^\intercal \boldsymbol{A}\boldsymbol{\theta} = \sum_{\ell \in P} \sum_{k \in P} a_{\ell,k}\theta_k\theta_\ell$ is *design compatible* with respect to a design and its set of unobserved pairs Ω if the probability of simultaneously observing θ_i and θ_j is zero only when the corresponding element in $\boldsymbol{A}$ is zero:

$$\forall k, \ell \in P,\ (k, \ell) \in \Omega \implies a_{k\ell} = 0,$$

where $a_{k\ell}$ is the element in the kth row and ℓth column of $\boldsymbol{A}$.

Proposition 3.3. *An unbiased estimator exists for a quadratic form if and only if it is design compatible.*

A seemingly simple way to make quadratic forms design compatible is to use a design for which $\Pr(k, \ell \in S)$ is not zero for any pair $k, \ell \in P$. However, such designs are not possible because a unit cannot simultaneously be assigned to the two exposures to be contrasted. That is, we always have $\Pr(k, \ell \in S) = 0$ whenever k and ℓ refer to two different potential outcomes for the same unit. This lack of observability is inescapable, prompting Holland (1986) to call this the fundamental problem of causal inference.

A possible remedy is to impose structural assumptions on the potential outcomes. Such assumptions allow us to extrapolate from observed to unobserved outcomes, which may allow for unbiased or consistent variance estimation. For example, Neyman (1923) notes that if the treatment effects are constant between units, the variance is estimable. However, these assumptions are generally strong, and rarely tenable in practice.

The variance estimation problem is exacerbated by interference and complex experimental designs. In interference settings, the structure of the exposure mapping often prevents certain pairs of exposures between different units to be inherently simultaneously unrealizable. For example, when units interact with each other in a network, all neighbors to a unit that is treated will necessarily be indirectly exposed to treatment. Therefore, it is impossible for two neighboring units to simultaneously receive the direct treatment and pure control exposures. In no-interference settings, some designs set $\Pr(k, \ell \in S)$ to zero, or a small value, even for pairs $k, \ell \in P$ that are in principle simultaneously observable. This could either be in an effort to improve the efficiency of the point estimator, such as in a matched-pair design, or because the design is forced on the experimenter by external factors.

3.2.5 Conservative variance bounds

After realizing that the variance of linear estimators cannot be estimated well, experimenters often opt for a conservative variance estimator. That is, they accept an estimator that systematically overestimates the variance, providing a pessimistic assessment of the precision of the point estimator. An inference procedure, such as

a confidence interval, based on a conservative variance estimator errs on the side of caution, in the sense that it motivates firm conclusions only under disproportionately strong evidence in favor of the conclusion.

We may understand such conservative variance estimators as estimators of an upper bound on the variance. We say that a function $\mathrm{VB}\colon \mathbb{R}^{2n} \to \mathbb{R}$ that satisfies $\mathrm{VB}(\boldsymbol{\theta}) \geq \boldsymbol{\theta}^\intercal \boldsymbol{A}\boldsymbol{\theta} = \mathrm{Var}(\widehat{\tau})$ for all potential outcomes $\boldsymbol{\theta} \in \mathbb{R}^{2n}$ is a *variance bound*. If we construct the function so it complies with the structure of simultaneous observability of the potential outcomes, we can construct an estimator for $\mathrm{VB}(\boldsymbol{\theta})$ corresponding to the true potential outcome vector $\boldsymbol{\theta}$, thereby yielding a conservative estimator of the variance.

The focus in this chapter is when the upper bound itself is a positive semidefinite quadratic form:

$$\mathrm{VB}(\boldsymbol{\theta}) = \boldsymbol{\theta}^\intercal \boldsymbol{B}\boldsymbol{\theta} = \sum_{k=1}^{2n}\sum_{\ell=1}^{2n} b_{k\ell}\theta_k\theta_\ell,$$

where $\boldsymbol{B}$ is a $2n$-by-$2n$ positive semidefinite matrix, and $b_{k\ell}$ is the element in the kth row and ℓth column of $\boldsymbol{B}$. Throughout the remainder of the chapter, we will use $\boldsymbol{B}$ to refer to both the coefficient matrix and the variance bound function $\mathrm{VB}(\boldsymbol{\theta})$.

To serve as a basis for a conservative estimator, we require the variance bounds to be both conservative and design compatible. This imposes two types of constraints on the coefficient matrix $\boldsymbol{B}$. To satisfy design conservativeness, $\boldsymbol{B}$ must be larger than $\boldsymbol{A}$ in the sense that $\boldsymbol{\theta}^\intercal \boldsymbol{A}\boldsymbol{\theta} \leq \boldsymbol{\theta}^\intercal \boldsymbol{B}\boldsymbol{\theta}$ for all vectors $\boldsymbol{\theta}$. This is precisely the *Loewner partial ordering* on symmetric matrices where $\boldsymbol{A} \preceq \boldsymbol{B}$ if $\boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite. To satisfy design compatibility, $\boldsymbol{B}$ must be such that $b_{k\ell} = 0$ for all pairs $(k, \ell) \in \Omega$. We refer to symmetric matrices that satisfy these two conditions as valid variance bounds.

Definition 3.4. A symmetric matrix $\boldsymbol{B}$ is a *valid* variance bound for $\boldsymbol{A}$ if it is larger than $\boldsymbol{A}$ in the Loewner order and design compatible under the current design. Let $\mathcal{B}$ collect all valid variance bounds:

$$\mathcal{B} = \big\{\boldsymbol{B} : \boldsymbol{A} \preceq \boldsymbol{B} \text{ and } b_{k\ell} = 0 \text{ for all } (k, \ell) \in \Omega\big\}.$$

An alternative, but equivalent, way to characterize the set of variance bounds is to use a slack matrix $\boldsymbol{S}$. A variance bound is constructed by adding the slack matrix to the variance matrix: $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{S}$. The resulting variance bound is conservative if and only if $\boldsymbol{S} = \boldsymbol{B} - \boldsymbol{A}$ is positive semidefinite. Thus, the slack captures what we add to the variance matrix in order to attain design compatibility. The set of slack matrices that produces valid variance bounds is

$$\mathcal{S} = \{\boldsymbol{S} : \boldsymbol{S} \succeq \boldsymbol{0} \text{ and } s_{k\ell} = -a_{k\ell} \text{ for all } (k, \ell) \in \Omega\},$$

where $s_{k\ell}$ is the element in the kth row and ℓth column of $\boldsymbol{S}$. We can reproduce the set of valid variance bounds as $\mathcal{B} = \{\boldsymbol{A} + \boldsymbol{S} : \boldsymbol{S} \in \mathcal{S}\}$. While the two representations

are equivalent, it is often more convenient to work with slack matrices.

3.2.6 Examples from the previous literature

We can use the representation outlined in this section to understand existing variance estimators. Our first example is the variance estimator described by Neyman (1923) for the difference-in-means estimator under a complete randomization design under no interference. A complete randomization design selects uniformly at random a fixed number of the units, typically half of the total number of units, to be assigned treatment. Because there is no interference, the relevant exposures are $a = 1$ and $b = 0$, and the estimand is the conventional average treatment effect.

Neyman (1923) showed that the variance in this setting is

$$\mathrm{Var}(\widehat{\tau}) = \frac{1}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i(1) - \frac{1}{n}\sum_{j=1}^{n} y_j(1)\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(y_i(0) - \frac{1}{n}\sum_{j=1}^{n} y_j(0)\right)^2 + 2\rho^2\right],$$

where ρ^2 in the third term is

$$\rho^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i(1) - \frac{1}{n}\sum_{j=1}^{n} y_j(1)\right)\left(y_i(0) - \frac{1}{n}\sum_{j=1}^{n} y_j(0)\right).$$

The first two terms are the population variances of the two potential outcomes, and the third term is their covariance. It is the covariance that is not estimable. Neyman's solution is to use the Cauchy–Schwarz inequality followed by the AM-GM inequality on ρ^2 to obtain the upper bound

$$\mathrm{Var}(\widehat{\tau}) \leq \frac{2}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\left(y_i(1) - \frac{1}{n}\sum_{j=1}^{n} y_j(1)\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left(y_i(0) - \frac{1}{n}\sum_{j=1}^{n} y_j(0)\right)^2\right] .$$

This upper bound can be estimated by the corresponding sample variances. Imbens and Rubin (2015) provide a more thorough treatment of this variance estimator.

We now show how the Neyman bound can be written using the framework we have described in this section. Using the linear coefficients corresponding to the difference-in-menas estimator, as discussed in Section 3.2.2, the variance of the estimator can be written $\mathrm{Var}(\widehat{\tau}) = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{A}\boldsymbol{\theta}$, where the covariance matrix of the linear coefficient vector is

$$\boldsymbol{A} = \mathrm{Cov}(\boldsymbol{v}) = \frac{1}{n(n-1)}\begin{bmatrix}\boldsymbol{H} & \boldsymbol{H} \\ \boldsymbol{H} & \boldsymbol{H}\end{bmatrix} \qquad \text{where} \qquad \boldsymbol{H} = \boldsymbol{I} - \mathbf{1}\mathbf{1}^\mathsf{T}/n,$$

and the potential outcome vector is $\boldsymbol{\theta} = \big(y_1(1), \ldots y_n(1), y_1(0), \ldots y_n(0)\big)$.

The matrix $\boldsymbol{A}$ is not design compatible because the diagonal elements in the off-diagonal blocks are non-zero, but the corresponding pairs of potential outcomes are never simultaneously observed. For example, $(1, n+1) \in \Omega$, so the product

$\theta_1\theta_{n+1} = y_1(1)y_1(0)$ is never observed, but entry in row 1 and column $n+1$ in $\boldsymbol{A}$ is non-zero. To address this, the Neyman estimator implicitly uses the slack matrix

$$\boldsymbol{S} = \frac{1}{n(n-1)}\begin{bmatrix} \boldsymbol{H} & -\boldsymbol{H} \\ -\boldsymbol{H} & \boldsymbol{H} \end{bmatrix}, \text{ yielding the variance bound } \boldsymbol{B} = \frac{2}{n(n-1)}\begin{bmatrix} \boldsymbol{H} & 0 \\ 0 & \boldsymbol{H} \end{bmatrix}.$$

This bound is a valid because $\boldsymbol{A} \preceq \boldsymbol{B}$, and $b_{k\ell} = 0$ for all $(k, \ell) \in \Omega$.

Our second example is the class of variance estimators described by Aronow and Samii (2013, 2017). The authors consider variance estimation for the Horvitz–Thompson point estimator under arbitrary experimental designs, and they describe a bound based on Young's inequality for products. The most straightforward version of Young's inequality states that $2ab \leq a^2 + b^2$ for any two real numbers a and b. It is possible to use this inequality to construct variance estimators for all linear point estimators.

The representation of the estimator provided by Aronow & Samii is somewhat unwieldy, so we refer interested readers to their papers. Instead, we provide the quadratic form representation of their variance estimator using the framework described in this section. For a variance

$$\mathrm{Var}(\widehat{\tau}) = \boldsymbol{\theta}^\intercal \boldsymbol{A}\boldsymbol{\theta} = \sum_{k=1}^{2n}\sum_{\ell=1}^{2n} a_{k\ell}\theta_k\theta_\ell,$$

with terms $(k, \ell) \in \Omega$ containing unobservable products $\theta_k\theta_\ell$, Aronow & Samii apply Young's inequality on each term separately:

$$a_{k\ell}\theta_k\theta_\ell \leq \frac{|a_{k\ell}|}{2}\big(\theta_k^2 + \theta_\ell^2\big).$$

We can write these inequalities as slack matrix. Let $\boldsymbol{M}_{k\ell}$ be a $2n \times 2n$ matrix with zeros entries except in the (k, ℓ)th block, which instead is given by

$$\begin{array}{c} \\ k \\ \ell \end{array}\begin{array}{c} \begin{array}{cc} k & \ell \end{array} \\ \begin{pmatrix} |a_{k\ell}| & -a_{k\ell} \\ -a_{k\ell} & |a_{k\ell}| \end{pmatrix} \end{array}.$$

The variance bound $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{S}$ underlying the Aronow & Samii variance estimator is given by the slack matrix

$$\boldsymbol{S} = \frac{1}{2}\sum_{(k,\ell)\in\Omega} \boldsymbol{M}_{k\ell}.$$

This bound is design compatible because $s_{k\ell} = -a_{k\ell}$ by construction for all $(k, \ell) \in \Omega$. Furthermore, because all matrices $\boldsymbol{M}_{k\ell}$ for $(k, \ell) \in \Omega$ are positive semidefinite, so will their sum be. Hence, $\boldsymbol{S}$ is positive semidefinite, and the bound is conservative.

3.3 Constructing Variance Bounds

3.3.1 Admissibility

Some valid variance bounds $\boldsymbol{B} \in \mathcal{B}$ will introduce slack beyond what is required for design compatibility, meaning that they are unnecessarily conservative. Experimenters will typically want to use a variance bound that introduces as little conservativeness, or slack, as possible. The amount of slack introduced will depend on the potential outcomes, so there is no universal ordering of the bounds with respect to conservativeness. This means that there exists no universally best bound. But some bounds can be ruled out because they introduce more slack than some other bound no matter what the potential outcomes might be. The following notion of inadmissibility characterizes such bounds.

Definition 3.5. A variance bound $\boldsymbol{B} \in \mathcal{B}$ is *inadmissible* if there exists another bound $\tilde{\boldsymbol{B}} \in \mathcal{B}$ such that $\boldsymbol{\theta}^\mathsf{T}\tilde{\boldsymbol{B}}\boldsymbol{\theta} \leq \boldsymbol{\theta}^\mathsf{T}\boldsymbol{B}\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \mathbb{R}^{2n}$ and $\boldsymbol{\theta}^\mathsf{T}\tilde{\boldsymbol{B}}\boldsymbol{\theta} < \boldsymbol{\theta}^\mathsf{T}\boldsymbol{B}\boldsymbol{\theta}$ for at least one $\boldsymbol{\theta} \in \mathbb{R}^{2n}$. Equivalently, $\boldsymbol{B}$ is inadmissible if there exists another bound $\tilde{\boldsymbol{B}}$ such that the difference $\boldsymbol{B} - \tilde{\boldsymbol{B}}$ is positive semidefinite and not zero. A variance bound that is not inadmissible is said to be *admissible*.

The set of admissible bounds are the minimal elements of $\mathcal{B}$ with respect to the Loewner order. Because the Loewner order is a partial order, there will generally be many minimal elements, mirroring the fact that no universally best bound exists. Moreover, there will generally be infinitely many admissible variance bounds. The focus in this section is how one should pick one of these admissible bounds to use in the construction of a variance estimator.

Proposition 3.6. *The Aronow–Samii bound is inadmissible in general experimental settings.*

Proof. A simple example suffices to prove the proposition. Indeed, the illustration in Section 3.1.1 proves the proposition. Observe that the variance is encoded by the matrix

$$\boldsymbol{A} = \frac{1}{4}\begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}.$$

The two bounds in the illustration are

$$\boldsymbol{B}_1 = \frac{1}{4}\begin{bmatrix} 2 & 0 & 0 & -2 \\ 0 & 2 & -2 & 0 \\ 0 & -2 & 2 & 0 \\ -2 & 0 & 0 & 2 \end{bmatrix} \quad \text{and} \quad \boldsymbol{B}_2 = \frac{1}{4}\begin{bmatrix} 3 & 0 & 0 & -1 \\ 0 & 3 & -1 & 0 \\ 0 & -1 & 3 & 0 \\ -1 & 0 & 0 & 3 \end{bmatrix},$$

of which $\boldsymbol{B}_2$ is the Aronow–Samii bound. Both these bounds are conservative and

design compatible, so they valid according to Definition 3.4. However, the difference

$$\boldsymbol{B}_2 - \boldsymbol{B}_1 = \frac{1}{4}\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

is positive semidefinite, and not zero, so Definition 3.5 tells us that $\boldsymbol{B}_2$ is inadmissible. □

It is possible to construct more involved experimental settings (e.g. larger samples, more intricate designs) where the Aronow–Samii bound is inadmissible, but we omit those examples in the interest of space. There are experimental settings where the Aronow–Samii bound is admissible. In Appendix B.3, we show that the Aronow–Samii bound is admissible in the no-interference setting when all pairs of assignments are observed with nonzero probability. Even so, the Aronow–Samii bound will generally be inadmissible under interference or complex designs.

3.3.2 Variance bound programs

There is currently no method for producing an admissible variance bound for general exposure mappings and designs. We propose a computational approach, where a variance bound from $\mathcal{B}$ is selected using an optimization formulation. For some real-valued function g on symmetric matrices, we aim to find a slack matrix $\boldsymbol{S} \in \mathcal{S}$ that minimizes g. Once an optimal $\boldsymbol{S}$ is obtained, we construct the variance bound as $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{S}$. Effectively, the choice of the objective function g implicitly selects the variance bound. We refer to this procedure as OPT-VB, which is formally described in Algorithm 2.

Algorithm 2: OPT-VB

Input: Objective function g

1 Compute a slack matrix $\boldsymbol{S}$ by solving the optimization program

$$\boldsymbol{S}^* \in \operatorname*{arg\,min}_{\boldsymbol{S} \in \mathcal{S}} g(\boldsymbol{S}). \tag{OPT}$$

2 Construct variance bound as $\boldsymbol{B} \leftarrow \boldsymbol{A} + \boldsymbol{S}$.

3 **return** *variance bound* $\boldsymbol{B}$.

The key aspect of OPT-VB is selecting the objective function g. The ideal objective function when our goal is to minimize the conservativeness is $g(\boldsymbol{S}) = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{S}\boldsymbol{\theta}$ for the true potential outcomes. But such an objective requires exact knowledge of the potential outcomes, making it infeasible. Our suggestion is instead to have g encode preferences of the experimenter concerning trade-offs associated with the bound, and any background knowledge they may have.

Independently of preferences and background knowledge, however, we always want to pick a bound that is admissible. The following definition and proposition provide conditions on the objective function g that ensure that the variance bound produced by OPT-VB is admissible.

Definition 3.7. A real-valued function f on symmetric matrices is *strictly monotone* if $f(\boldsymbol{A}) < f(\boldsymbol{A} + \boldsymbol{S})$ for all symmetric $\boldsymbol{A}$ and nonzero positive semidefinite $\boldsymbol{S}$.

Theorem 3.8. *If g is strictly monotone, then* OPT-VB *returns a variance bound that is conservative, design compatible and admissible.*

Using Theorem 3.8, we can verify that a variance bound obtained by OPT-VB is admissible by showing that the objective is strictly monotone. Indeed, unless otherwise noted, all objective functions discussed in this chapter are strictly monotone according to Definition 3.7, so Theorem 3.8 guarantees admissibility of the resulting bound. The proof of Theorem 3.8 appears in Appendix B.2.2.

Our definition of strict monotonicity differs from the conventional definition based on the strict Loewner order. This definition states that a strictly monotone function f satisfies $f(\boldsymbol{A}) < f(\boldsymbol{B})$ whenever $\boldsymbol{A} \prec \boldsymbol{B}$. The conventional definition, though well-motivated in many applications, does not align with our notion of admissibility, motivating the variation we present in Definition 3.7.

Theorem 3.8 motivates us to pick an objective that is strictly monotone no matter our preferences and background knowledge. Another concern that is universal is the computational tractability of solving the optimization problem underlying OPT-VB. As a rule of thumb, optimization is tractable if it is a convex program (Rockafellar, 1993). The set of slack matrices $\mathcal{S}$ is convex. Therefore, the optimization underlying OPT-VB is convex if g is selected to be a convex function. All objective functions discussed in this chapter are convex, so they admit efficient algorithms for finding an optimal solution, up to desired tolerance. We will not review convex programming in this chapter, but a wide variety of general purpose solvers are available (see, e.g., Dunning et al., 2017; Udell et al., 2014). Boyd and Vandenberghe (2004) provides an introduction to the theory of convex programming.

3.3.3 Norm objectives

We will first consider the situation in which the experimenter has little or no background knowledge about the potential outcomes. In this setting, the objective function encodes the experimenter's risk preference concerning the bound. In particular, at heart of the problem is a trade-off between average performance and worst-case performance of variance bounds. One may select the bound to not introduce much conservativeness for most potential outcomes, in a sense that will be made formal shortly, but then the bound can be excessively conservative for some potential outcomes. Conversely, one may select the bound to never be excessively conservative, but then it will be more conservative on average. A risk tolerant experimenter would

prefer the first type of bound, while a risk averse experimenter would prefer the second type.

We will use the family of Schatten p-norms for matrices to make this idea precise. Informally, our goal is to select a variance bound that is not excessively conservative as a function of the potential outcomes. A way to measure the magnitude of a quadratic form is by a matrix norm of its coefficient matrix. Thus, to make the variance bound small, we select a coefficient matrix $\boldsymbol{B}$. We use Schatten norms because they allow us to capture the risk trade-off.

To understand the Schatten norm, we must consider the spectral decomposition of the coefficient matrices $\boldsymbol{B} \in \mathcal{B}$ of all valid bounds. This decomposition allows us to write the variance bound as

$$\mathrm{VB}(\boldsymbol{\theta}) = \boldsymbol{\theta}^\intercal \boldsymbol{B} \boldsymbol{\theta} = \|\boldsymbol{\theta}\|^2 \sum_{k=1}^{2n} w_k \lambda_k,$$

where $w_k = \langle \boldsymbol{\eta}_k, \boldsymbol{\theta} \rangle^2 / \|\boldsymbol{\theta}\|^2$ captures the alignment of the potential outcome vector to the kth eigenvector $\boldsymbol{\eta}_k$ of $\boldsymbol{B}$, and λ_k is the corresponding kth eigenvalue. Because $\boldsymbol{B}$ is positive semidefinite, all eigenvalues are non-negative. By construction, the coefficients w_k are non-negative and sum to one, so they act as weights in a convex combination of the eigenvalues. That is, the conservativeness of the variance bound is determined by the eigenvalues and the alignment of the potential outcomes to the eigenvectors of $\boldsymbol{B}$. Thus, if we make the eigenvalues of the variance bound matrix $\boldsymbol{B}$ small, we ensure that the bound is not excessively conservative.

The Schatten norms are different ways of measuring the magnitude of the eigenvalues. Formally, a Schatten p-norm of $\boldsymbol{B}$ is the usual p-norm applied to the vector of singular values of $\boldsymbol{B}$, which in our case coincide with the eigenvalues:

$$\|\boldsymbol{B}\|_p = \left(\sum_{k=1}^{2n} |\lambda_k|^p \right)^{1/p}.$$

The Schatten p-norm thus acts in the same way as vector p-norm: as p becomes larger, the norm becomes disporportionally affected by large eigenvalues. If p is small, a few eigenvalues are allowed to be large if it means that many other eigenvalues can be small. But, if p is large, the focus is mainly on making the largest eigenvalues small, at the cost of increasing smaller eigenvalues. Therefore, a risk averse experimenter would want to use a Schatten p-norm with a large p, ensuring that no eigenvalue is much larger than the others. A risk tolerant experimenter would instead prefer a Schatten p-norm with a smaller p, as this will ensure that the sum of eigenvalues is not too large. The following proposition shows that the resulting bound is admissible no matter the choice of p.

Proposition 3.9. *For all $p \in [1, \infty)$, the Schatten p-norm objective $g(\boldsymbol{S}) = \|\boldsymbol{A} + \boldsymbol{S}\|_p$ is strictly monotone, ensuring that the variance bound produced by* OPT-VB *using g*

is admissible.

The Schatten p-norm coincides with some more familiar matrix norms for particular values of p. If we set $p = 1$, the Schatten p-norm is simply the sum of the absolute value of the eigenvalues. This coindices with the nuclear norm, which also is called the trace norm, commonly defined as $\|\boldsymbol{B}\|_1 = \mathrm{tr}(\sqrt{\boldsymbol{B}^\intercal \boldsymbol{B}}) = \mathrm{tr}(\boldsymbol{B})$, where the last equality holds for symmetric and positive semidefinite $\boldsymbol{B}$. Using this norm provides a bound with the best average performance, in the sense that it puts uniform weight on all eigenvalues no matter their magnitude. An experimenter with complete tolerance for risk would pick this norm to use for the objective function.

At the other extreme, if we let $p \to \infty$, we obtain the operator norm induced by the 2-norm. This norm is more commonly defined as $\|\boldsymbol{B}\|_\infty = \max_{\|\boldsymbol{\theta}\|_2=1} \|\boldsymbol{B}\boldsymbol{\theta}\|_2$. When $\boldsymbol{B}$ is positive semidefinite, as in our case, the norm is equal to the maximum eigenvalue of $\boldsymbol{B}$. An experimenter who is maximally risk adverse would pick this norm to use for the objective function, as it would trade-off any amount of average conservativeness for even a minimal reduction is worst-case conservativeness. The spectral norm is not strictly monotone according to Definition 3.7, so an arbitrary minimizer of an objective function using this norm is not guaranteed to be admissible. However, experimenters are rarely so risk averse that they would use the spectral norm, and objectives based on the Schatten p-norm for any $p < \infty$ is strictly monotone. Nevertheless, we sketch an argument in Section 3.3.6 which shows that there always exists an admissible minimizer of the spectral norm.

The final special case we consider is when $p = 2$. This recovers the Frobenius norm, which more commonly is defined as $\|\boldsymbol{B}\|_2 = \sqrt{\mathrm{tr}(\boldsymbol{B}\boldsymbol{B}^\intercal)}$. As made clear by the connection to the Schatten norm, the Frobenius norm can also be interpreted as the Euclidean norm of the eigenvalues. Hence, the Frobenius norm provides an intermediate point in the risk trade-off; it penalizes large eigenvalues disproportionately much, making sure that no eigenvalue gets very large, but it does not ignore the smaller eigenvalues completely. We believe many experimenters will find that the Frobenius norm is an appropriate choice in many setting. But, of course, the choice should ultimately be governed by each experimenter's risk preference in the particular application.

3.3.4 Targeted linear objectives

The class of norm objectives can encode experimenters' risk preferences, but they cannot encode background knowledge they might have about the potential outcomes. We describe a class of *targeted* objective functions to fill this role. The choice of a particular objective function from this class will reflect the experimenter's prior substantive knowledge. An experimenter's prior knowledge about the potential outcomes need not be correct in any way to ensure the validity of the resulting variance bound, but if the experimenter is able to provide accurate information about the potential outcomes, the bound will be less conservative. This idea is related to the

model-assisted tradition used in design-based survey sampling (Särndal et al., 1992).

The class of *targeted linear objectives* takes the form

$$g(\boldsymbol{S}) = \langle \boldsymbol{S}, \boldsymbol{W} \rangle,$$

where $\boldsymbol{S}$ is the slack matrix to be evaluated, $\boldsymbol{W}$ is a targeting matrix of the same dimensions, and $\langle \cdot, \cdot \rangle$ denotes the trace inner product on matrices, defined formally as

$$\langle \boldsymbol{S}, \boldsymbol{W} \rangle = \operatorname{tr}(\boldsymbol{S}\boldsymbol{W}) = \sum_{i,j} s_{i,j} w_{i,j},$$

where $s_{i,j}$ and $w_{i,j}$ are elements of $\boldsymbol{S}$ and $\boldsymbol{W}$, respectively. As we discuss to some detail in the next section, $\boldsymbol{W}$ is used to target particular potential outcomes, perhaps motivated by prior substantive knowledge. Importantly, all objective functions in this class are linear in the coefficients of the slack matrix, so the optimization problem underlying OPT-VB is a semidefinite program when used by these targeted linear objectives, ensuring computational tractability.

By construction, the bound returned by OPT-VB using a targeted linear objective will be valid. What makes the class of targeted linear objectives stand out compared to the norm objectives is a type of completeness result. The class of targeted linear objectives characterizes the set of all admissible variance bounds.

Theorem 3.10. *A variance bound $\boldsymbol{B}$ is admissible if and only if can be obtained from* OPT-VB *using the objective function $g(\boldsymbol{S}) = \langle \boldsymbol{S}, \boldsymbol{W} \rangle$ for some positive definite targeting matrix $\boldsymbol{W}$.*

The proof that every bound returned by OPT-VB using a positive definite targeting matrix is admissible proceeds by showing that every targeted linear objective is strictly monotone and then appeals to Theorem 3.8. The proof of the opposite direction, that every admissible bound can be obtained as a solution OPT-VB using some targeted linear objective is more involved and appeals to the separating hyperplane theorem from convex analysis.

Theorem 3.10 has several implications. First, it shows that we always obtain an admissible when we use a targeted linear objective with a positive definite targeting matrix. Second, it allows us to re-interpret other procedures for constructing variance bounds by showing what matrix they implicitly target, which by extension shows what potential outcomes they implicitly targets. We explore this in Appendix B.3, where we show that the Aronow–Samii bound may be obtained for certain designs in the no-interference setting by using a diagonal targeting matrix $\boldsymbol{W}$. However, it might be challenging to derive the targeting matrix for a given admissible variance bound and so this type of re-interpretation is possible only when the targeting matrix is known beforehand.

3.3.5 Choosing targeting matrices

There are many ways to construct the targeting matrix $\boldsymbol{W}$ to use with a targeted linear objective. It is beyond the scope of this paper to fully investigate all possible approaches, but we give some suggestions here.

We begin with an example to illustrate the underlying idea. Recall that the value of the variance bound using coefficients $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{S}$ is

$$\mathrm{VB}(\boldsymbol{\theta}) = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{B}\boldsymbol{\theta} = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{\theta}^\mathsf{T}\boldsymbol{S}\boldsymbol{\theta}.$$

On the right hand side, the first term is the true variance, so it does not depend on the bound. The second term is the excess conservativeness introduced to make the bound design compatible, so it does depend on the bound, as shown by the inclusion of the slack matrix. Suppose that the experimenter's prior substantive knowledge suggests, perhaps incorrectly, that the potential outcomes are close to some vector $\boldsymbol{v} \in \mathbb{R}^{2n}$. If $\boldsymbol{v}$ in fact is the true potential outcomes $\boldsymbol{\theta}$, this second term is known, so it can be used directly to select the bound. That is, the experimenter is here motivated to use the objective function $g(\boldsymbol{S}) = \boldsymbol{v}^\mathsf{T}\boldsymbol{S}\boldsymbol{v}$.

If $\boldsymbol{v}$ is the true potential outcomes, the variance bound obtained using $g(\boldsymbol{S}) = \boldsymbol{v}^\mathsf{T}\boldsymbol{S}\boldsymbol{v}$ as objective has the minimal amount of slack required for design compatibility. If $\boldsymbol{v}$ is similar but not exactly the same as the true potential outcomes, the bound will often perform well, provided that the slack matrix is sufficiently well behaved to the extent that the resulting quadratic form is sufficiently smooth. If $\boldsymbol{v}$ is very different from the true potential outcomes, the bound is still valid, but it could be excessively conservative. However, importantly for our purposes, this is a targeted linear objective with targeting matrix $\boldsymbol{W} = \boldsymbol{v}\boldsymbol{v}^\mathsf{T}$, because $\boldsymbol{v}^\mathsf{T}\boldsymbol{S}\boldsymbol{v}$ can be written as $\langle \boldsymbol{S}, \boldsymbol{v}\boldsymbol{v}^\mathsf{T} \rangle$.

It is rare that experimenters have so precise background knowledge so they can produce a single vector $\boldsymbol{v}$ of potential outcomes to target. And even if they could, it would generally not be advisable to do so. The matrix $\boldsymbol{W} = \boldsymbol{v}\boldsymbol{v}^\mathsf{T}$ will not be full rank, so it is not positive definite. This means that the produced bound may be inadmissible, and it could be excessively conservative if the targeted vector be very different from the true potential outcomes. To address this, we will explore targeting several potential outcome vectors simultaneously in the rest of this section.

Consider a situation where the experimenter suspect that the true potential outcomes are similar to at least one vector in a collection of m vectors: $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$. A natural objective function here is

$$g(\boldsymbol{S}) = \sum_{i=1}^{m} q_i \boldsymbol{v}_i^\mathsf{T}\boldsymbol{S}\boldsymbol{v}_i,$$

where $q_i \geq 0$ is some weight of the ith vector $\boldsymbol{v}_i$, indicating its importance or relevance.

This is also a targeted linear objective, corresponding to the targeting matrix

$$\boldsymbol{W} = \sum_{i=1}^{m} q_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal.$$

If the true potential outcomes are similar to one or more vectors in $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$, the produced bound can be expected to perform well. Including many vectors in $\boldsymbol{W}$ will make the targeting less sharp, potentially making the bound more conservative even if the true potential outcome is similar to one of the targeted vectors. While this suggests that fewer vectors should be targeted, experimenters should generally be motivated to include many vectors. As noted above, it is generally advisable to ensure that $\boldsymbol{W}$, and that requires that there are at least $2n$ linearly independent vectors that are targeted, but they need not be targeted to the same degree. A simple way to achieve this is to add a scaled identity matrix to a sum of a handful of vectors. That is, for some m much smaller than $2n$ and some small $\gamma > 0$, the following objective is positive definite:

$$\boldsymbol{W} = \sum_{i=1}^{m} q_i \boldsymbol{v}_i \boldsymbol{v}_i^\intercal + \gamma \boldsymbol{I}.$$

A convenient and more general way to express these weighted averages of potential outcome vectors is as a generative model. That is, we would consider $\boldsymbol{\theta}$ as a random variable drawn from some distribution. To replicate the objective with m vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$, the random variable $\boldsymbol{\theta}$ would take the value $\boldsymbol{v}_i$ with probability $q_i / \sum_i^m q_i$. The targeting matrix is then produced by taking the expectation of the outer product of the random variable: $\boldsymbol{W} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal]$. We use the subscript $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ to indicate that the expectation is taken over the imagined distribution of $\boldsymbol{\theta}$, rather than the true randomization distribution induced by the experimental design, as in the rest of the paper. Seen from this perspective, the targeted linear objective minimizes the expected variance bound:

$$\mathbb{E}_{\boldsymbol{\theta}}[\mathrm{VB}(\boldsymbol{\theta})] = \left\langle \boldsymbol{B}, \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal] \right\rangle = \langle \boldsymbol{A}, \boldsymbol{W} \rangle + \langle \boldsymbol{S}, \boldsymbol{W} \rangle,$$

where as above, only $\left\langle \boldsymbol{S}, \boldsymbol{W} \right\rangle$ depends on the choice of the bound.

It should be emphasized here that the interpretation of $\boldsymbol{\theta}$ as a random variable is simply a convenient way to express large collections of potential outcome vectors. It is not assumed nor required for any of our results that whatever distribution experimenters choose to use here accurately reflect how the potential outcome actually was generated; the resulting bound is always valid, and it is admissible provided that $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal]$ is positive definite. However, the bound will be less conservative if the distribution is a good approximation to the true potential outcomes.

One advantage with expressing the targeting matrix as $\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal]$ for some imagined distribution is that practitioners are often more comfortable expressing background

knowledge they might have in this form rather than explicit collections of vectors. One example uses covariates to inform the choice of the targeting matrix.

Consider a situation where the experimenter believes the potential outcomes can be well-approximated by a linear function of the covariates. That is, we can write the true potential outcomes as

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{X} & 0 \\ 0 & \boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_a \\ \boldsymbol{\beta}_b \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_a \\ \boldsymbol{\varepsilon}_b \end{bmatrix},$$

for some small vectors $\boldsymbol{\varepsilon}_a$ and $\boldsymbol{\varepsilon}_b$. We write the linear function in this form because we do not want to use the same linear function to approximate both types of potential outcomes that are stacked in $\boldsymbol{\theta}$. The matrix $\boldsymbol{X}$ is observed and fixed, but we can represent our partial ignorance about $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \boldsymbol{\varepsilon}_a, \boldsymbol{\varepsilon}_b)$ as a distribution. There are many ways of doing this, including using pilot or other related studies. For illustration here, we will consider all them as normally distributed. In that case, the corresponding targeting matrix becomes

$$\boldsymbol{W} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal] = \begin{bmatrix} \boldsymbol{X}\boldsymbol{X}^\intercal + \sigma^2 \boldsymbol{I} & 0 \\ 0 & \boldsymbol{X}\boldsymbol{X}^\intercal + \sigma^2 \boldsymbol{I} \end{bmatrix},$$

where σ^2 denotes the relative variance of an element in the coefficient vectors $\boldsymbol{\beta}_d$ and an element in $\boldsymbol{\varepsilon}_d$. This objective is positive definite for all $\sigma > 0$.

More intricate generative working models may be used to construct more elaborate targeting matrices. One may choose to work with more sophisticated generative working models, such as those based on the underlying exposure mapping in a social network experiment (Basse et al., 2016; Toulis and Kao, 2013).

3.3.6 Composite objectives

There are situations where experimenters want a combination of the properties offered by the objective functions discussed in this section. Using the fact that monotonicity is maintained under positive combinations, the following proposition shows that any combination of elementary objectives can be used with OPT-VB.

Proposition 3.11. *If a set of m functions $g_1, \ldots, g_m$ are monotone and at least one of the functions are strictly monotone, then for any set of positive coefficients $\gamma_1, \ldots, \gamma_m$, the function $g_c = \sum_{i=1}^m \gamma_i g_i$ is strictly monotone. Thus,* OPT-VB *returns a variance bound that is conservative, design compatible and admissible when called with the composite objective g_c.*

One example where a composite objective would be useful is when an experimenter who has detailed background knowledge still wants to control the worst-case conservativeness of the bound. As described in Sections 3.3.4 and 3.3.5, the experimenter would encode their knowledge in a targeting matrix $\boldsymbol{W}$ to be used with a targeted linear objective. But the experimenter might worry that their perceived knowledge

is inaccurate, in which case the produced bound risks being excessively conservative, which especially is the case if $\boldsymbol{W}$ is not full rank. To address this, the experimenter could use a composite objective that includes the operator norm, as discussed in Section 3.3.3. Recall that the operator norm captures the worst-case conservativeness of the bound. For some coefficient γ, deciding the relative focus between targeting and the worst-case outcome, the resulting composite objective is

$$g(\boldsymbol{S}) = \langle \boldsymbol{S}, \boldsymbol{W} \rangle + \gamma \|\boldsymbol{A} + \boldsymbol{S}\|_\infty.$$

If $\boldsymbol{W}$ is positive definite, this composite objective is strictly monotone, so it yields a bound that is conservative, design compatible and admissible.

Another example when a composite objective is useful is when an experimenter is interested in minimizing worst-case conservativeness alone, but still wants to ensure that the bound they use is admissible. Recall from Section 3.3.3 that the operator norm is monotone, but not strictly monotone, so it could yield inadmissible bounds. This could be addressed by using a Schatten p-norm for a very large p, but such a solution would suffer from numerical instability. An alternative is to regularize the operator norm with the Frobenius norm in a composite objective. That is, for some small $\gamma > 0$, one would use the objective

$$g(\boldsymbol{S}) = \|\boldsymbol{A} + \boldsymbol{S}\|_\infty + \gamma \|\boldsymbol{A} + \boldsymbol{S}\|_2^2.$$

Proposition 3.11 applies here because the operator norm is monotone and the Frobenius norm is strictly monotone, meaning that the composite objective is strictly monotone and yields a bound that is conservative, design compatible and admissible. One would want to set γ to as small value as possible here, as this will ensure that the worst-case conservativeness indeed is minimized as well as possible. It would be possible study the solution in the limit $\gamma \to 0$, but we omit that here in the interest of space.

3.4 Testing Admissibility of Variance Bounds

In Section 3.3, we introduced the notion of an admissibility and proposed several methods for computing admissible variance bounds. In this section, we present a semidefinite program for testing admissibility of a given variance bound. This allows experimenters to test admissibility of a given variance bound—obtained by possibly different methods than those presented above—before running an experiment.

The procedure Test-Admissibilitydecides whether a variance bound is admissible by testing whether the optimal value of a particular semidefinite program is positive. Recall that a variance bound $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{S}$ is inadmissible if there exists another variance bound $\tilde{\boldsymbol{B}} = \boldsymbol{A} + \tilde{\boldsymbol{S}}$ such that

$$\boldsymbol{B} - \tilde{\boldsymbol{B}} = \boldsymbol{S} - \tilde{\boldsymbol{S}}$$

is nonzero and positive semidefinite; on the other hand, if we can certify that no such matrix exists, then the variance bound $\boldsymbol{B}$ is admissible. To this end, the procedure Test-Admissibility searches over all slack matrices $\tilde{\boldsymbol{S}} \in \mathcal{S}$ with the extra constraint that $\boldsymbol{S} - \tilde{\boldsymbol{S}}$ is positive semidefinite. What remains to be shown then, is whether there exists a feasible solution such that this difference is nonzero. To determine this, Test-Admissibility maximizes the trace of the difference, $\operatorname{tr}(\boldsymbol{S} - \tilde{\boldsymbol{S}})$. If the optimal value is positive, then the difference is nonzero and the original variance bound is inadmissible; otherwise, the optimal value is zero and the variance bound is admissible. The test for admissibility is given formally below in Algorithm 3.

Algorithm 3: Test-Admissibility

Input: Variance bound slack matrix $\boldsymbol{S}$ and unobservable pairs Ω

1 Solve the following semidefinite program

$$\begin{aligned} \alpha \leftarrow \underset{\tilde{\boldsymbol{S}}}{\text{maximize}} \quad & \operatorname{tr}(\boldsymbol{S} - \tilde{\boldsymbol{S}}) \\ \text{subject to} \quad & \tilde{x}_{i,j} = s_{ij} \ \text{ for all } (i,j) \in \Omega, \\ & 0 \preceq \tilde{\boldsymbol{S}} \preceq \boldsymbol{S}. \end{aligned} \qquad \text{(Admissible-SDP)}$$

2 **return** `False` *if optimal value* $\alpha > 0$ *and* `True` *otherwise.*

Note that $\tilde{\boldsymbol{S}} = \boldsymbol{S}$ is always a feasible solution to the optimization underlying Test-Admissibility, but this yields an objective value of zero. The following theorem guarantees correctness of the Test-Admissibility procedure.

Theorem 3.12. Test-Admissibility *returns* `True` *if and only if the variance bound is admissible.*

Proof. Suppose that the variance bound $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{S}$ is admissible. Then, there does not exist a matrix $\tilde{\boldsymbol{S}} \in \mathcal{S}$ such that $\boldsymbol{S} - \tilde{\boldsymbol{S}}$ is positive semidefinite and nonzero. Thus, the only feasible solution to (Admissible-SDP) is $\boldsymbol{S}$, which yields an objective value of 0. In this case, Test-Admissibility returns `True`, which is the correct answer.

Suppose that the variance bound is inadmissible. Then, there exists a matrix $\tilde{\boldsymbol{S}} \in \mathcal{S}$ such that $\boldsymbol{S} - \tilde{\boldsymbol{S}}$ is positive semidefinite and nonzero. This matrix $\tilde{\boldsymbol{S}}$ is feasible and yields a positive objective value, ie. $\operatorname{tr}(\boldsymbol{S} - \tilde{\boldsymbol{S}}) > 0$. This inequality follows because the trace is the sum of the eigenvalues and the matrix $\boldsymbol{S} - \tilde{\boldsymbol{S}}$ has non-negative eigenvalues and at least one positive eigenvalue, as it is positive semidefinite and nonzero. In this case, Test-Admissibility returns `False`, which is the correct answer. □

There are some numerical considerations when implementing Test-Admissibility in practice. Namely, semidefinite programs can only be solved up to some desired accuracy. This means that testing whether the optimal objective is exactly zero is generally not possible using finite precision arithmetic, except in certain restricted cases. For this reason, the main practical use case of Test-Admissibility will be

to certify that a variance bound is sufficiently inadmissible, rather than certifying admissibility. This numerical issue should not be of great concern, as an experimenter can use Test-Admissibility to certify that a variance bound is approximately admissible (up to an arbitrary desired tolerance) which is generally suitable in practice.

In order to decide that the input variance bound is inadmissible, Test-Admissibility needs only to produce a feasible solution where $\operatorname{tr}(\boldsymbol{S} - \tilde{\boldsymbol{S}}) > 0$. If the input variance bound is inadmissible, this may require significantly less computation time than solving the underlying optimization program to optimality. In this way, early stopping may be used in Test-Admissibility for increased computational efficiency.

3.5 Estimation of Variance Bounds

To construct an estimator of a variance bound, we use the fact that a quadratic form can be reinterpreted as a linear function of the elements $\theta_i\theta_j$ of the outer product $\boldsymbol{\theta}\boldsymbol{\theta}^\intercal$. For example, a Horvitz–Thompson estimator of a variance bound $\boldsymbol{B}$ is

$$\widehat{\mathrm{VB}}(\boldsymbol{\theta}) = \sum_{i \in S} \sum_{j \in S} \frac{b_{ij}\theta_i\theta_j}{\Pr(i, j \in S)}.$$

Any linear estimator can in principle be used to estimate the variance bound, but they will not perform equally well. It is beyond the scope of this work to investigate which of the estimators in the linear class is best used for variance bounds, and we restrict our discussion of estimation of the variance bound to the Horvitz–Thompson estimator above. We direct interested readers to Middleton (2020), who provides an in-depth discussion about estimators of quadratic forms that are design compatible.

When the variance bound is design-compatible, then the Horvitz–Thompson estimator will be unbiased; however, unbiasedness is a relatively weak condition and experimenters will typically want the estimator to also have high precision. The precision of the estimator depends not only on the variance bound itself, but also on the underlying design and the potential outcomes. To this end, for each pair of outcomes $(i, j) \in [2n] \times [2n]$, define the inverse propensity indicator variable to be

$$R_{(i,j)} = \frac{\mathbf{1}[i, j \in S]}{\Pr(i, j \in S)}$$

and collect these n^2 random variables into a vector, denoted $\boldsymbol{R}_{\bar{\Omega}}$. Let $\operatorname{Cov}(\boldsymbol{R}_{\bar{\Omega}})$ be the covariance matrix of the inverse propensity indicator vector. The following theorem presents a finite sample bound on the mean squared error of the Horvitz–Thompson estimator for the variance bound.

Proposition 3.13. *Suppose that the variance bound $\boldsymbol{B}$ is design-compatible. Then, the mean squared error of the Horvitz–Thompson estimator may be bounded as*

$$\mathbb{E}[(\,VB(\boldsymbol{\theta}) - \widehat{VB}(\boldsymbol{\theta}))^2] \leq \|\operatorname{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_* \cdot \|\boldsymbol{B}\|_F^2 \cdot M^2 \ ,$$

where $M = \max_{i \in [2n]} |\theta_i|$ *is the largest absolute value of all potential outcomes.*

Proposition 3.13 upper bounds the mean squared error of the Horvitz–Thompson estimator of the variance bound into the product of three distinct terms: one corresponding to the design, one corresponding to the variance bound, and one corresponding to the potential outcomes. The term $\|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_*$ measures how well-behaved the design is. This term may be large when either many second order exposure probabilities $\Pr(i, j \in S)$ are very small or when pairs of exposures are highly correlated across many units. Consistent estimation of the variance bound is impossible under either of these conditions. In the no-interference setting, we have that $\|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_* = 1$ under a Bernoulli design and more generally, $\|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_* = B$ under an independent cluster design, where B is the size of the largest cluster. Ultimately, the term $\|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_*$ can only be rigorously investigated for a fixed exposure mapping and design; however, we suspect that under mild interference and accommodating designs, this term can be treated as a constant.

The terms $\|\boldsymbol{B}\|_F$ and M in Proposition 3.13 measure the magnitude of the coefficients in the variance bound and the potential outcomes, respectively. When these terms are large, the magnitude of the variance bound itself can becomes large, thereby increasing the mean squared error. In other words, Proposition 3.13 shows that the Horvitz–Thompson estimator achieves higher precision when the design is well-behaved, the coefficients in the variance bound are not execssively large, and the potential outcomes are bounded.

As a corollary, we obtain conditions under which the Horvitz–Thompson estimator is a consistent estimator of the variance bound.

Corollary 3.14. *Suppose that the variance bound* $\boldsymbol{B}$ *is design compatible and that the terms* $\|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_*$ *and* M *are constant in an asymptotic sequence. If* $\|\boldsymbol{B}\|_F^2 \to 0$ *in the asymptotic sequence, then the Horvitz–Thompson estimator is a consistent estimator of the variance bound:* $\mathbb{E}[(VB(\boldsymbol{\theta}) - \widehat{VB}(\boldsymbol{\theta}))^2] \to 0$.

Proposition 3.13 and Corollary 3.14 place a bound on the absolute value of potential outcomes. In Appendix B.2.3, we show how that this may be replaced by a weaker bound on fourth moments of the potential outcomes, although this changes the norm used to measure the magnitude of the coefficients in the variance bound.

Corollary 3.14 demonstrates that in order to consistently estimate the variance bound, the coefficients of the bound must be shrinking in the asymptotic sequence. For consistent linear estimators of the treatment effect, one should expect that the coefficients of the variance matrix $\boldsymbol{A}$ are decreasing. For example, in the no-interference setting, the Horvitz–Thompson estimator under the Bernoulli design satisfies $\|\boldsymbol{A}\|_F^2 = 2/n$, which goes to zero. Heuristically speaking, precise estimation of the variance bound is possible when it is chosen so as not to substantially increase the squared Frobenius norm.

In light of Proposition 3.13 and Corollary 3.14, experimenters may wish to choose the objective in OPT-VB so that the resulting variance bound may be estimated

with high precision with a Horvitz–Thompson estimator presented above. To this extent, we propose the following regularized objective that encourages such variance bounds:

$$g(\boldsymbol{S}) = \langle \boldsymbol{S}, \boldsymbol{W} \rangle^2 + \alpha \|\boldsymbol{A} + \boldsymbol{S}\|_F^2 \ ,$$

where $\alpha \geq 0$ is a penalty parameter set by the experimenter. The first term aims to produce a variance bound with small slack in the targeted potential outcome directions, while the second term acts as penalty that prevents the variance bound from having excessively large coefficients. The penalty parameter α facilitates this trade-off.

We may further motivate the regularized objective in terms of the bias-variance trade-off of the variance bound estimator to the true (unknown) variance. We write $\mathrm{V}(\boldsymbol{\theta})$ to denote the variance of the linear estimator, as a function of the potential outcomes. As an estimator of the variance, the estimator features the following bias-variance decomposition:

$$\mathbb{E}[(\mathrm{V}(\boldsymbol{\theta}) - \widehat{\mathrm{VB}}(\boldsymbol{\theta}))^2] = (\mathrm{VB}(\boldsymbol{\theta}) - \mathrm{V}(\boldsymbol{\theta}))^2 + \mathbb{E}[(\mathrm{VB}(\boldsymbol{\theta}) - \widehat{\mathrm{VB}}(\boldsymbol{\theta}))^2] \ .$$

The first term on the right hand side is the square of the slack introduced in the variance bound for outcome $\boldsymbol{\theta}$. The second term is the mean squared error of the variance bound estimator to the variance bound. Using the trace formulation of the slack term and the upper bound of the MSE in Proposition 3.13, we have that the MSE of the estimator to the true variance is at most

$$\mathbb{E}[(\mathrm{V}(\boldsymbol{\theta}) - \widehat{\mathrm{VB}}(\boldsymbol{\theta}))^2] \leq \langle \boldsymbol{S}, \boldsymbol{\theta}\boldsymbol{\theta}^\mathsf{T} \rangle^2 + \Big(\|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_* \cdot M^2 \Big) \|\boldsymbol{B}\|_F^2 \ .$$

The regularized objective can be understood as this upper bound on the MSE. Rather than using a specific outcome vector $\boldsymbol{\theta}$, we use a positive semidefinite matrix $\boldsymbol{W}$ in order to target a smaller bias across many potential outcomes, as discussed in Section 3.2.5 Using the upper bound above, we may consider setting the penalty term as $\alpha = \|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_* \cdot M^2$, when we have some idea about the magnitudes of the potential outcomes.

3.6 Conclusion and Open Problems

In this chapter, we have presented methodology for variance estimation under interference and complex designs. We have characterized the variance of linear estimators and derived the form of admissible upper bounds. We presented OPT-VB, an optimization-based procedure that enables experimenters to select an admissible variance bound based on their risk aversion and prior substantive knowledge. Once selected, the bound itself may be estimated using any arbitrary linear estimator. As discussed in Section 3.5, the choice of variance estimator may influence the selection of the bound itself.

Properties of the variance bound and its estimator depend on the design and the estimator of the treatment effect. In this chapter, we have considered the design and estimator to be fixed, but they are ultimately chosen by the experimenter. Thus, we recommend that the experimenter choose the design, effect estimator, variance bound, and variance bound estimator together in order to achieve a more holistic approach to experimental design. Such a holistic approach may be guided by the result of a pilot study, a model of a data-generating process, or a worst-case analysis.

There are several open problems suggested by this work, which we list below.

- **Closed Form Variance Bounds**: *For specific experimental settings, derive closed form admissible variance bounds.* In this work, we derive a computational procedure for deriving variance bounds under an arbitrary exposure mapping, design, and (linear) treatment effect estimator. The benefit of such a computational technique is its wide applicability, but the downside is that it adds a layer of complexity which may discourage certain experimenters from adopting this approach. Deriving closed form variance bounds for specific experimental settings would provide more transparent (if less flexible) methods of variance estimation.

- **Beyond Quadratic Variance Bounds**: *Extend the variance bound selection methodology beyond quadratic functions.* In this work, we considered only quadratic upper bounds. Although the class of quadratic functions was natural for a variety of reasons, several proposed variance estimators Aronow et al. (2014); Menzel and Imbens (2021) rely on variance bounds which are not quadratic forms. These variance estimators, however, are only applicable in restricted settings and the way in which they resolve the implicit trade-off in variance bounds is unclear. Combining these approaches would be of great methodological interest.

Chapter 4

Bipartite Experiments Under a Linear Exposure-Response Assumption [1]

In previous chapters, we considered two problems—covariate balancing in Chapter 2 and variance estimation in Chapter 3—which arise in the traditional potential outcomes framework for designing and analyzing randomized experiments. In this chapter, we go beyond the traditional setting and study the bipartite experimental framework, where units that receive treatment are distinct from units on which outcomes are measured. Our main contributions are new estimation and inference methodology for treatment effects in the bipartite experimental framework, under a structural assumption on the outcomes.

4.1 Introduction

Two-sided marketplaces are rife with interesting but difficult causal questions. What happens to demand if shipping times or fees are reduced? What happens to people's willingness to use ride-hailing apps if more drivers are enrolled in specific cities? What happens to long term user behavior if a hotel booking platform changes its recommendation engine? The causal impact of these changes is hard to quantify, even when using randomized experiments, because marketplace dynamics often violate a central tenet of conventional experimentation: the Stable Unit Treatment Value Assumption, abbreviated SUTVA. This assumption stipulates that the treatment assigned to one unit does not affect any other units. Violations of this assumption is a phenomenon known as interference, which is often present in the case of marketplace experiments and complicates causal analysis.

The bipartite experimental framework offers a useful formalism to study two-sided market experiments and other violations of SUTVA that can happen along the edges of a bipartite graph. This stands in contrast with interference that occurs on graphs

[1] Based on the working paper: Christopher Harshaw, Fredrik Sävje, David Eisenstat, Vahab Mirrokni, and Jean Pouget-Abadie (2021) "Design and Analysis of Bipartite Experiments under a Linear Exposure-Response Model". arXiv:2103.06392.

where all units are of the same type (e.g. users of a social network). In the bipartite experimental framework, we distinguish two types of units: units that can be subject to an intervention and units whose responses are of interest to the experimenter. We assign treatment to the former and measure the outcomes of the latter. The causal impact of treating one group of units is measured on the other group by tracking the *exposure* to treatment that the latter group receives, informed by the knowledge of the bipartite graph between them. We remark that the treatment status of a single unit may affect the measured outcomes of many units and, likewise, a measured outcome may be affected by many treatment units.

For example, consider a marketplace where buyers compete for limited goods, some of which may be perfectly or partially substitutable. Their demand of these goods form a bipartite graph that potentially can be inferred by the marketplace owner. The owner of the marketplace would like to determine the causal effect of discounting prices on buyers' marketplace behavior through a randomized experiment. Randomly assigning certain buyers to receive a discounted price is often not possible, and might even be prohibited, in which case randomization is only possible at the item-level. At the same time, simply comparing discounted goods with non-discounted goods runs the risk of severe bias: a discounted good may do well against a non-discounted substitutable good, which does not accurately reflect a world where either both or neither are discounted. Instead, the marketplace owner decides to monitor this change at the buyer level, positing that, by tracking both their behavioral changes and their exposure to discounted goods, the causal effect of the discount can be teased out.

As is done in much of the interference literature and other settings where SUTVA is violated, assumptions on potential outcomes are made when the bipartite graph has a many-to-many structure in order to allow for tractable inference. One such assumption is existence of an exposure mapping, which posits that outcomes are some simple function of the treatment assignments of neighboring units in the bipartite graph (Toulis and Kao, 2013; Aronow and Samii, 2017). In this work, we study estimation of an all-or-nothing treatment effect in the bipartite experimental framework under a linear exposure-response model, where exposures are linear functions of assignments and responses are linear functions of the exposures. The main contributions of this chapter are summarized as follows:

- We describe the Exposure-Reweighted Linear (ERL) estimator, an unbiased linear estimator of the average total treatment effect under the linear exposure-response model. We show that the ERL estimator is consistent and asymptotically normal, provided the graph remains sufficiently sparse.

- We describe a variance estimator, which may be used to construct confidence intervals via a normal approximation. We show that under mild conditions on the exposure distribution, the variance estimator is unbiased. We achieve unbiasedness without assuming constant treatment effects or any other restrictions on the heterogeneity between units' potential outcomes.

- We describe EXPOSURE-DESIGN, a cluster-based design which aims to increase the precision of the ERL estimator. The design achieves this by increasing the variance of individual exposures while decreasing the covariance between different exposures. This improves precision is several settings of interest.

4.1.1 Related works

Within the wide-ranging causal inference literature, our work falls squarely within the subset relying on the potential outcomes framework (Neyman, 1923; Imbens and Rubin, 2015). The design and analysis of randomized experiments in the presence of interference has garnered plenty of attention, spanning vaccination trials (Struchiner et al., 1990), agricultural studies (Kempton, 1997), voter-mobilization field experiments (Sinclair et al., 2012), and viral marketing campaigns (Aral and Walker, 2011; Eckles et al., 2016b). It is beyond the scope of this chapter to extensively review the literature causal inference with interference. Instead, we direct readers to the review article by Halloran and Hudgens (2016).

Our work is primarily motivated by marketplace experiments. Evidence of interference in marketplaces has been noted across industries for various experimental designs (Gupta et al., 2019). Reiley (2006), Einav et al. (2011) and Holtz et al. (2020) study the interference bias that results from supply-side randomization, while Blake and Coey (2014) and Fradkin (2015) consider this problem in the context of demand/user-side randomization. Basse et al. (2016) and Liu et al. (2020) compare supply-side randomization to two-sided randomization as well as to budget-split designs, showing bias can be reduced in the context of certain ad auction experiments. More recently, Johari et al. (2020) characterize which randomization scheme (supply-side, demand-side, or two-sided) leads to reduced bias as a function of market balance.

We consider a slightly different experimental setting, introduced by Zigler and Papadogeorgou (2021), characterized by random assignment of treatment on one side of the bipartite graph (demand- or supply-side), while outcomes are measured on the other side. The advantage of this framework is that complex interference relationship can be captured by an exposure function (similar to Aronow and Samii (2017)), which is assumed to solely determine an unit's outcome. This makes the problem of estimating causal effects tractable despite the complex interference structure. Zigler and Papadogeorgou (2021) study causal estimands which are more closely related to direct effects rather than the all-or-nothing treatment effect considered here. In addition, the analysis of their estimators requires that the bipartite graph be the union of small connected components.

Motivated by marketplace experiments, Pouget-Abadie et al. (2019) introduce a cluster-based design for general bipartite graphs in this framework and consider a similar estimand and exposure-response assumption. Later, Doudchenko et al. (2020) proposed a class of generalized propensity score estimators for this framework, which are unbiased for both experimental and observational settings under standard assumptions and a similar exposure-response assumption.

Our work is the first to propose methods for provably valid inference (e.g., confidence intervals) in the bipartite settings and to jointly consider estimators and designs which improve overall precision of treatment effect estimators. While the cluster design of Pouget-Abadie et al. (2019) is based on the intuition of achieving a large spread of exposures, it disregards the correlation of exposures and is not rigorously tied to the performance of an estimator. Additionally, while the estimators proposed by Doudchenko et al. (2020) are unbiased, they are based on a different approach which requires fitting a generalized propensity score function. Neither of these papers present methods for valid inference.

4.2 Experimental Setting

In the bipartite experimental framework, the units which receive treatment are distinct from the units on which the outcomes are measured. For example, Zigler and Papadogeorgou (2021) apply the framework to analyze how interventions on power plants' pollution affect the hospitalization rates among nearby hospitals. We discuss the general bipartite framework in Section 4.2.1 and the linear exposure response assumption in Section 4.2.2.

4.2.1 Bipartite experiments

In the bipartite experiment setting, there are two groups of units: the *diversion units*, to which treatment is applied, and the *outcome units*, where outcomes are measured. We denote the set of m diversion units by V_d and the set of n outcome units by V_o.

Each of the m diversion units receives a (random) binary treatment $z_i \in \{\pm 1\}$, and we collect these treatments into a treatment vector, $\boldsymbol{z} = (z_1, z_2, \dots z_m) \in \{\pm 1\}^m$. The distribution over the random treatment vectors is called the *design* of the experiment and it is chosen by the experimenter. Each of the outcome units $i \in V_o$ is associated with a potential outcome function $y_i(\boldsymbol{z})$, which maps the treatment assignments to the observed value, which is a real number. In the bipartite setting, we assume that each potential outcome function depends only on the treatment of a neighborhood set of diversion units. More formally, there exists a *neighborhood mapping* $\mathcal{N} : V_o \to 2^{V_d}$ such that for all outcome units $i \in V_o$,

$$y_i(\boldsymbol{z}) = y_i(\boldsymbol{z}') \quad \text{ if } z_j = z'_j \text{ for all } j \in \mathcal{N}(i) \ .$$

Throughout the chapter, we assume that the neighborhood mapping is known and correctly specified, so that the above condition holds. We recover the standard Stable Unit Treatment Value Assumption (SUTVA) when the diversion units are identified with the outcome units and the neighborhood mapping is the identity function.

The number of potential outcomes for each outcome unit grows exponentially in the size of its neighborhood. Zigler and Papadogeorgou (2021) avoid this issue by assuming that the bipartite structure is the union of many small connected components.

Unfortunately, this is typically not a reasonable assumption in the marketplace settings where we know that more varied interactions occur: buyers may interact with a variety of products. Without further restrictions on the structure of the neighborhoods or the potential outcome functions, inference of any causal estimand is impossible (Basse and Airoldi, 2018; Sävje et al., 2021). Take, for example, an instance where the neighborhood of each outcome unit is all diversion units. For this reason, we introduce a stronger assumption on the potential outcomes.

4.2.2 Linear exposure-response model

In order to admit tractable inference of causal estimands, we consider a linear exposure-response model, which consists of two underlying assumptions: a linear exposure assumption and a linear response assumption, which we state formally below.

In the linear exposure-response model, we suppose that there is a weighted bipartite graph between diversion units and outcomes units, where the edges have non-negative weights $w_{i,j} \geq 0$, which we arrange into an n-by-m incidence matrix $\boldsymbol{W}$. An edge $w_{i,j}$ represents the influence of diversion unit j on the outcome units i. We say that outcome unit i and diversion unit j are *incident* if the weight $w_{i,j}$ is positive. The degree of a diversion unit is the number of outcome units it is incident to, and the largest degree among all diversion units is denoted d_d. The degree of an outcome unit is defined similarly and the largest degree among all outcome units is denoted d_o. For simplicity, we assume that each outcome unit has degree at least 1 and the weights incident to an outcome unit are normalized to sum to one, i.e. the rows of the incidence matrix $\boldsymbol{W}$ sum to one. We also assume that this weighted bipartite graph is known to the experimenter. In many market experiments, the experimenter may construct an approximation of this graph from historical data.

The *linear exposure assumption* is that the treatment assignments influences the potential outcomes only through a linear combination, which is more structured than arbitrary influence. More formally, for each outcome unit $i \in V_o$, the *exposure* of outcome unit i is

$$x_i(\boldsymbol{z}) = \sum_{j \in V_d} w_{i,j} z_j \ ,$$

and for all pairs of assignment vectors $\boldsymbol{z}$ and $\boldsymbol{z}'$ with $x_i(\boldsymbol{z}) = x_i(\boldsymbol{z}')$, we have that $y_i(\boldsymbol{z}) = y_i(\boldsymbol{z}')$. This implies that the neighborhood mapping is such that $\mathcal{N}(i) = \{j : w_{i,j} > 0\}$.

We arrange these n exposures into an exposure vector $\boldsymbol{x}(\boldsymbol{z}) = (x_1(\boldsymbol{z}), x_2(\boldsymbol{z}) \dots x_n(\boldsymbol{z}))$. Because the exposure is a function of treatment, the experimental design completely determines the exposure distribution. This linear exposure assumption is a generalization of the partial and stratified interference assumptions discussed by Hudgens and Halloran (2008). When the treatment assignment vector $\boldsymbol{z}$ is clear from context, we write simply x_i and $\boldsymbol{x}$ for the ith exposure and the exposure vector, respectively. Using matrix-vector notation, we may write the exposure vector as $\boldsymbol{x}(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$. Due

to the normalization of the weights and the ± 1 values of the treatment assignment, each exposure takes values in the range $[-1, 1]$.

The *linear response assumption* is that for each outcome unit, the potential outcome is a linear function of its exposure. That is, for each outcome unit $i \in V_o$, there exists parameters α_i and β_i such that

$$y_i(\boldsymbol{z}) = \alpha_i + \beta_i x_i(\boldsymbol{z}) \ .$$

We refer to α_i as the unit-specific intercept and β_i as the unit-specific slope. These terms are unknown to the experimenter, and the experimenter only observes the outcome $y_i(\boldsymbol{z})$, along with the sampled assignment vector $\boldsymbol{z}$ and the resulting exposure vector $\boldsymbol{x}$.

We refer to the *linear exposure-response model* as the combination of the linear exposure assumption and the linear response assumption. The linear exposure-response model places certain limits on the potential outcomes, but allows for more complex structure in the bipartite graph. This trade-off is preferable in settings such as marketplace experiments, where we know that a complex bipartite structure exists and we are more comfortable with making simplifying assumptions about potential outcomes. For further discussion on empirical and theoretical evidence for complex structure in marketplace experiments, we refer the reader to Blake and Coey (2014); Fradkin (2017); Johari et al. (2020).

Structural assumptions on the outcomes similar to the linear exposure-response assumption presented here are commonly made throughout the interference literature. The *linear-in-means* (LIM) model posits that a unit's response is a linear function of their own treatment, and the mean of the treatments of their group (Manski, 1993). The LIM model has been extended in various ways in the context of partial interference (Baird et al., 2018; Offer-Westort and Dimmery, 2021) and social network experiments (Bramoullé et al., 2009; Toulis and Kao, 2013). Chin (2019b) investigates the use of machine learning estimators for the global average treatment effect under a variation of the LIM when the terms in the linear model of arbitrary functions of treatment. Basse et al. (2016) study model-assisted estimators and designs under the "normal sum-model" which is similar to the linear exposure-response considered here, but with a normal noise term. We remark that the bipartite setting with the linear exposure-response assumption recovers the standard SUTVA setting when diversion units are identified with the outcome units and the weight matrix is the identity.

From one perspective, the linear exposure-response model is a strong assumption. It requires that the response for each unit is exactly a linear function in the exposure. This rules out, for example, that different diversion units have different impacts on a single outcome unit. But from another perspective, the model is completely unrestrictive: it does not limit the heterogeneity between units at all. That is, knowing the response function for one unit tells us nothing about the response function of other units. While there are few settings in which the linear exposure-response model will hold exactly, it will often be a useful approximation given its unrestrictiveness

with respect to heterogeneity. In Section 4.5, we analyze the behavior of the ERL estimator under a general non-linear response assumption, finding that it estimates a best linear approximation to the average response. However, we leave it to future work to more finely characterize the behavior of estimator under general responses and we assume the linear exposure-response model holds exactly throughout the rest of the chapter.

4.2.3 Causal estimand

We are interested in understanding the contrast between two possible worlds: one where all diversion units receive treatment and one where they all receive control. For an individual outcome unit, this contrast is captured by the individual treatment effect, $\tau_i = y_i(\boldsymbol{z} = \mathbf{1}) - y_i(\boldsymbol{z} = -\mathbf{1})$ for $i \in V_o$. Just as in the typical SUTVA setting, we cannot hope to estimate the individual treatment effects well because only one potential outcome is observed for any one unit. In light of this, we opt to estimate an aggregated causal quantity. In this chapter, we are interested in the Average Total Treatment Effect (ATTE), which is the average contrast between the scenario that all diversion units receive treatment and all diversion units receive control. More precisely, ATTE is defined as

$$\tau = \frac{1}{n}\sum_{i=1}^{n} \tau_i = \frac{1}{n}\sum_{i=1}^{n}\Big[y_i(\boldsymbol{z} = \mathbf{1}) - y_i(\boldsymbol{z} = -\mathbf{1})\Big]$$

Under the linear exposure-response model, the ATTE is proportional to the average of the slope terms, as shown in the following proposition.

Proposition 4.1. *Under the linear exposure-response assumption, the ATTE is* $\tau = \frac{2}{n}\sum_{i=1}^{n}\beta_i$.

Proof. The individual treatment effect of outcome unit i is proportional to its slope, as

$$\tau_i = y_i(\boldsymbol{z} = \mathbf{1}) - y_i(\boldsymbol{z} = -\mathbf{1}) = \big[\beta_i x_i(\boldsymbol{z} = \mathbf{1}) + \alpha_i\big] - \big[\beta_i x_i(\boldsymbol{z} = -\mathbf{1}) + \alpha_i\big] = 2\beta_i \ ,$$

where we have used that $x_i(\boldsymbol{z} = \mathbf{1}) = 1$ and $x_i(\boldsymbol{z} = -\mathbf{1}) = -1$. The result follows by taking the average of the individual treatment effects. □

There are two main challenges in estimating the ATTE in this setting: we want to estimate the average of the slopes of many different linear response functions, but only one point from each of the distinct linear response functions is observed. Although stated in somewhat unfamiliar terms, this is the fundamental problem of causal inference (Holland, 1986). The second challenge is that of constructing a treatment design which realizes a desirable exposure distribution. As previously discussed at the end of Section 4.2.2, this is a difficult task when the bipartite weight matrix has

non-trivial overlapping structures. In the remainder of the chapter, we focus on addressing these two challenges by developing an estimator and a class of designs which together accurately estimate the ATTE.

4.2.4 Cluster designs

Some of the analysis in this chapter assumes that the treatment is assigned according to a *independent cluster designs*, where the diversion units are grouped into clusters and treatment is assigned to an entire cluster. More formally, we say that a partition $C_1, C_2, \ldots, C_\ell$ of the diversion units is a *clustering*, which we denote as $\mathcal{C} = \{C_1, C_2, \ldots, C_\ell\}$. That is, all clusters are disjoint and the union of all clusters is set of diversion units V_d. Given a clustering $\mathcal{C}$, a treatment assignment from the corresponding *independent cluster design* is drawn in the following way: independently for each cluster, we assign all diversion within a cluster to have either treatment $z_i = 1$ with probability p and treatment $z_i = -1$ with probability $1 - p$. For notational simplicity, we consider the treatment probability p to be fixed for all clusters, but our results extend to the setting where each cluster has its own treatment probability. Note that the class of independent cluster designs is completely specified by $\mathcal{C}$ and p.

4.3 The Exposure Reweighted Linear Estimator

We describe the Exposure Reweighted Linear Estimator, which is an unbiased estimate of the ATTE under the linear exposure-response assumption. The Exposure Reweighted Linear (ERL) estimator is defined below as

$$\hat{\tau} = \frac{2}{n} \sum_{i=1}^{n} y_i(\boldsymbol{z}) \left(\frac{x_i(\boldsymbol{z}) - \mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))} \right) . \tag{4.1}$$

The ERL estimator requires knowledge of the mean and variance of each of the marginal exposure distributions under the treatment design. For several commonly used designs such as Bernoulli and independent cluster designs, these exposure characteristics may be computed directly; however, for arbitrary designs, the expectation and variance of the exposures may need to be estimated to high precision using samples drawn from the treatment design. We assume here that these exposure characteristics are known exactly. We emphasize that the ERL estimator may be used under any treatment design and not just the cluster-based treatment design we propose in Section 4.6.

The ERL estimator belongs to the class of linear estimators, as it is a (random) linear function of the observed outcomes. It shares similarities with the style of Horvitz–Thompson estimators (Narain, 1951; Horvitz and Thompson, 1952), but is not the same. The Horvitz–Thompson estimators re-weights an outcome by the prob-

ability of observing that outcome, while the ERL estimator re-weights an outcome by the normalized distance of the exposure from its mean. When there are many exposures, such as under the linear exposure-response model, the type of re-weighting done by the Horvitz–Thompson estimator would lead to excessively large variance.

4.3.1 Statistical analysis of the ERL estimator

In this section, we analyze the behavior of the ERL estimator as a point estimator of the average total treatment effect (ATTE). First, we show that the ERL estimator is unbiased. Then we show consistency and asymptotic normality of the ERL estimator, provided that the bipartite graph is not too dense. Theorem 4.2 below ensures that under mild conditions on the treatment design, there is no systematic bias in the ERL estimator.

Theorem 4.2 (Unbiasedness)**.** *Suppose the design is such that each exposure has a positive variance. Under the linear response assumption, the* ERL *estimator is unbiased for the ATTE:* $\mathbb{E}[\hat{\tau}] = \tau$.

Next, we analyze the asymptotic behavior of the ERL estimator. In the asymptotic analysis, we suppose that there is a sequence of bipartite experiments, in which the number of units is growing to infinity. Strictly speaking, all quantities of the experiment such as the bipartite graph, the outcomes, the treatment design, etc, should be indexed by an integer N; however, we drop these subscripts for notational clarity.

We make two additional assumptions about the bipartite experiments in this asymptotic sequence. The first is that the potential outcomes are bounded. The second is that the design has limited dependence between treatment assignments.

Assumption 4.3 (Bounded Potential Outcomes)**.** The potential outcomes are bounded in absolute value $|y_i(\boldsymbol{z})| \leq M$, where M is a constant.

Assumption 4.4 (Design Conditions)**.** The treatments assignments are distributed according to an independent cluster design, where the probability of treatment p is bounded away from 0 and 1 by a constant in the asymptotic sequence. Additionally, the sizes of clusters are bounded by k, a constant in the asymptotic sequence.

Assumption 4.4 rules out certain classes of treatment designs, such as complete randomization (i.e. group balanced designs). While it may be possible to obtain similar asymptotic results under such designs, we limit our consideration to those satisfying Assumption 4.4. Under these assumptions, we prove that ERL is consistent when the bipartite graph is not too dense.

Theorem 4.5 (Consistency)**.** *Under Assumptions 4.3 and 4.4, and supposing that* $d_d d_o^3 = o(n)$ *in the asymptotic sequence, the* ERL *estimator converges in mean square to the ATTE:* $\lim_{N\to\infty} \mathbb{E}[(\hat{\tau} - \tau)^2] = 0$.

The main technical assumption that we require for consistency is that $d_d d_o^3 = o(n)$ in the asymptotic sequence, where we recall that d_d and d_o are the maximum degrees of the diversion and outcome units, respectively. Informally, this condition ensures that the bipartite graph is not too dense as it grows. While consistency may hold under a weaker condition for a particular design, some density assumption like this must be made. As an example, consider the complete bipartite graph where all outcome units receive the same exposure, in which case consistent estimation is impossible.

We now discuss a setting where this condition $d_d d_o^3 = o(n)$ holds. Suppose that each diversion unit has fixed degree d_d, which is a constant with respect to m and n. The average degree of an outcome unit is then $\bar{d_o} = d_d(m/n)$. Assuming that the maximum outcome degree d_o is within a constant factor of the average, this yields that the term $d_d d_o^3 = \mathcal{O}(d_d^4(m/n)^3)$. Using that the diversion degrees are constant, we get that this term is bounded by $o(n)$ if $m = o(n^{4/3})$. Thus, in graphs with constant diversion degrees where the edges are evenly distributed between outcome units, the hypothesis of Theorem 4.5 holds when m grows at a rate slower than $n^{4/3}$.

We next describe the asymptotic distribution of the estimator. In particular, we show that the ERL estimator converges in distribution to a normal distribution as the size of the bipartite experiment grows, provided that the graph remains sparse. This result is derived under the same asymptotic regime as above. In order to prove the central limit theorem, we require an additional assumption on the asymptotic sequence of bipartite experiments. Namely, we require that the variance of the ERL estimator decreases no faster than the parametric rate.

Assumption 4.6 (Not Superefficient)**.** The normalized variance of the ERL estimator $n \operatorname{Var}(\hat{\tau})$ is bounded away from zero asymptotically.

Assumption 4.6 rules out settings in which we can estimate the ATTE at a faster than parametric rate. Such settings are theoretically possible, but not practically relevant. In particular, Assumption 4.6 rules out two scenarios. The first is when the magnitude of the potential outcomes approaches zero in the asymptotic sequence. Note that this requires that almost all potential outcomes approach zero; the magnitude of the potential outcomes are generally non-zero even when their average is zero. The second scenario is when the design close to perfectly pinpoints the potential outcomes. This can be formalized as the variance of each individual term of the estimator diminishes asymptotically, i.e. $\operatorname{Var}(\hat{\tau}_i) \to 0$, where $\hat{\tau}_i = 2y_i(\boldsymbol{z})(x_i - \mathbb{E}[x_i])/\operatorname{Var}(x_i)$. Both of these scenarios are knife-edge cases that we have good reason to believe would not materialize in practice. Even if they do, the estimator would still be unbiased and consistent, but its asymptotic distribution might not be normal.

We are now ready to present a central limit theorem which states that under mild regularity conditions, the ERL estimator is asymptotically normal.

Theorem 4.7 (Asymptotic Normality)**.** *Under Assumptions 4.3, 4.4, and 4.6, and*

supposing that $d_d^{1.6} d_o^4 = o(n)$*, the* ERL *estimator is asymptotically normal:*

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

The proof relies on Stein's method for bounding distances between distributions (see, e.g. Ross (2011)). We use Stein's method because standard techniques for establishing central limit theorems which heavily rely on independence are not applicable in the bipartite experimental framework where exposures are necessarily correlated. We remark that Stein's method has been recently used for obtaining limiting behavior of other estimators in the interference literature (Aronow and Samii, 2017; Chin, 2019a; Ogburn et al., 2020).

The assumptions on the asymptotic growth of the bipartite graph may be interpreted similarly as those appearing in Theorem 4.5. Namely, they prevent the bipartite graph from becoming too dense. We remark that the growth assumptions required for asymptotic normality (Theorem 4.7) are stronger than those required required for consistency (Theorem 4.5). The growth assumptions in Theorem 4.7 are only sufficient and we conjecture that they are not necessary for asymptotic normality. However, weakening these growth conditions would require a different analysis, either by a more careful application of Stein's method or by different means all together.

Assumption 4.4 allows for a broad class of designs. For example, unit-level Bernoulli randomization falls into this class, but this design does not at all consider the structure of the bipartite graph and will generally perform poorly. To derive analytical results for this broad class of designs, the growth conditions on the bipartite graph are quite restrictive, and they may be too restrictive in certain settings where more dense interaction patterns occur. If one restricts focus to a smaller class of designs, these growth conditions could potentially be weakened. The key implication of Assumption 4.4 together with the growth conditions is that the variance of the exposures is large and the correlation between most pairs of exposures is small. Heuristically, these conditions on the exposure distribution are the main aspects required for consistency and normality. We describe a design in Section 4.6 that directly targets the exposure distribution to satisfy these conditions, and it will therefore be better behaved than many of the designs allowed by Assumption 4.4.

4.4 Variance Estimation

In this section, we present methods for constructing confidence intervals for the ATTE in the bipartite setting under the linear exposure-response assumption. If we knew the variance of the ERL estimator, we could use Theorem 4.7 directly to construct asymptotically valid confidence intervals. Because the variance of the ERL estimator depends on the unobserved potential outcomes, we must construct an estimator of the variance.

In the typical experimental settings with SUTVA and binary treatments, unbiased variance estimation is not possible without strong assumptions on the heterogeneity between units (Imbens and Rubin, 2015). In light of this negative result, experimenters tend to favor conservative variance estimators that lead to valid but overly wide confidence intervals. In contrast to the typical SUTVA setting with binary treatments, we show that unbiased variance estimation is possible in the bipartite setting under the linear response assumption when the exposures take many (i.e. non-binary) values.

Our approach to constructing a variance estimator is to decompose the ERL estimator into a weighted average of individual effect estimators, and to decompose the variance of the ERL estimator as the average of the variances and covariances of these individual effect estimators. To this end, define $\hat{\tau}_i \triangleq 2y_i(\boldsymbol{z})(x_i - \mathbb{E}[x_i])/\operatorname{Var}(x_i)$ to be the individual terms in the ERL estimator. We may interpret $\hat{\tau}_i$ as an unbiased, but very imprecise, estimator of the individual treatment effect τ_i. The ERL estimator can be written as the average of these quantities: $\hat{\tau} = (1/n)\sum_{i=1}^n \hat{\tau}_i$. The variance of the ERL estimator may be decomposed as

$$\operatorname{Var}(\hat{\tau}) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^n \hat{\tau}_i\right) = \frac{1}{n^2}\sum_{i=1}^n \left[\operatorname{Var}(\hat{\tau}_i) + \sum_{j\neq i}\operatorname{Cov}(\hat{\tau}_i, \hat{\tau}_j)\right] \ .$$

We will construct unbiased estimators of $\operatorname{Var}(\hat{\tau}_i)$ and $\operatorname{Cov}(\hat{\tau}_i, \hat{\tau}_j)$ and average over these estimators to obtain an estimator of the variance of the ERL estimator. Because we are averaging over many unbiased estimators, the average will often be well-behaved even if the individual estimators perform poorly in mean squared sense.

We begin by deriving an estimator for the individual variance terms, $\operatorname{Var}(\hat{\tau}_i)$. To this end, we define the random variable

$$Q_i = \frac{(x_i - \mathbb{E}[x_i])^2}{\operatorname{Var}(x_i)^2} - \frac{\operatorname{Var}(x_i)(x_i^2 - \mathbb{E}[x_i^2]) - \operatorname{Cov}(x_i, x_i^2)(x_i - \mathbb{E}[x_i])}{\operatorname{Var}(x_i)\operatorname{Var}(x_i^2) - \operatorname{Cov}(x_i, x_i^2)^2} \ ,$$

which is a quadratic function of the exposure x_i. Because the exposure distribution is known to the experimenter, Q_i is an observable quantity. The following lemma demonstrates that by re-weighting the observed quantity $y_i(\boldsymbol{z})^2$ by Q_i, we obtain an unbiased estimate of the individual variance terms.

Lemma 4.8. *Fix an outcome unit $i \in V_o$. If the exposure x_i takes at least three values with non-zero probability, then the variance of unit i's individual treatment effect estimator is equal to*

$$\operatorname{Var}(\hat{\tau}_i) = 4 \cdot \mathbb{E}[y_i(\boldsymbol{z})^2 Q_i] \ .$$

We remark that the condition that the support of the exposure contains at least three points is critical for Lemma 4.8 to go through; in fact, when the exposure takes only binary values, then previously established results in the SUTVA setting

show that these individual variance terms cannot generally be estimated without bias (Imbens and Rubin, 2015). Moreover, because the exposure distribution is known before the experiment is run, the experimenter may determine whether this condition holds. At the end of this section, we discuss variance estimation when this support condition on the exposures does not hold.

We estimate the covariance terms $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$ in a similar manner. For each pair of outcome units $i, j \in V_o$, we define the random variable

$$R_{i,j} = \frac{(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j])}{\mathrm{Var}(x_i)\,\mathrm{Var}(x_j)} - \frac{a_{i,j}(x_i x_j - \mathbb{E}[x_i x_j]) + b_{i,j}(x_i - \mathbb{E}[x_i]) + c_{i,j}(x_j - \mathbb{E}[x_j])}{\Psi_{i,j}} \ ,$$

which is a bivariate degree two polynomial in the exposures x_i, x_j where the coefficients $a_{i,j}$, $b_{i,j}$, $c_{i,j}$, and $\Psi_{i,j}$ depend on the joint distribution of the pair of exposures. More precisely, these coefficients are defined as

$$\begin{aligned}
a_{i,j} &= \mathrm{Var}(x_i)\,\mathrm{Var}(x_j) - \mathrm{Cov}(x_i, x_j)^2, \\
b_{i,j} &= \mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_j) - \mathrm{Var}(x_j)\,\mathrm{Cov}(x_i x_j, x_i), \\
c_{i,j} &= \mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_i) - \mathrm{Var}(x_i)\,\mathrm{Cov}(x_i x_j, x_j), \\
\Psi_{i,j} &= \mathrm{Var}(x_i x_j)(\mathrm{Var}(x_i)\,\mathrm{Var}(x_j) - \mathrm{Cov}(x_i, x_j)^2) - \mathrm{Var}(x_i)\,\mathrm{Cov}(x_i x_j, x_j)^2 \\
&\quad - \mathrm{Var}(x_j)\,\mathrm{Cov}(x_i x_j, x_i)^2 + 2\,\mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_j)\,\mathrm{Cov}(x_i x_j, x_i) \ .
\end{aligned}$$

We remark that these coefficients depend only on the joint distribution of pairs of exposures and so they are known to the experimenter. Thus, just as Q_i is an observable quantity, so too is $R_{i,j}$. The following lemma demonstrates that reweighting the product of two potential outcomes $y_i(\boldsymbol{z})y_j(\boldsymbol{z})$ by $R_{i,j}$ yields an unbiased estimate of the individual covariance terms.

Lemma 4.9. *Fix a pair of outcome units $i \neq j \in V_o$. If $\Psi_{i,j} \neq 0$, then the covariance between individual treatment effect estimates $\hat{\tau}_i$ and $\hat{\tau}_j$ may be expressed as*

$$\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = 4 \cdot \mathbb{E}[y_i(\boldsymbol{z})y_j(\boldsymbol{z})R_{i,j}] \ .$$

The condition that $\Psi_{i,j}$ is nonzero ensures that the re-weighting factor $R_{i,j}$ is well-defined. One situation in which $\Psi_{i,j}$ is zero is when $x_i = x_j$ with probability one: the condition that $\Psi_{i,j}$ is nonzero rules out this scenario. We have not been able to construct a joint distribution for which the denominator $\Psi_{i,j}$ is zero without perfectly correlated exposures. We conjecture that no such distribution exists, but we have not been able to prove this yet. Regardless, the exposure distribution is known, so the experimenter can determine whether $\Psi_{i,j} = 0$ for any pairs of outcome units. At the end of this section, we discuss how one can proceed in the case that $\Psi_{i,j} = 0$ for some pairs of units.

Lemmas 4.8 and 4.9 together suggest the following variance estimator:

$$\widehat{\mathrm{Var}}(\hat{\tau}) \triangleq \frac{4}{n^2}\sum_{i=1}^{n}\Big[y_i(\boldsymbol{z})^2 Q_i + \sum_{j\neq i} y_i(\boldsymbol{z})y_j(\boldsymbol{z})R_{i,j}\Big] \ .$$

When the exposure distributions which are symmetric about 0 (e.g. an independent cluster design with treatment probability $p = 1/2$), the random variables Q_i and $R_{i,j}$ simplify considerably as the $\mathrm{Cov}(x_i x_j, x_i)$ and $\mathrm{Cov}(x_i x_j, x_j)$ terms are zero. In this case, these functions may be written simply as

$$Q_i = \frac{x_i^2}{\mathrm{Var}(x_i)^2} - \frac{x_i^2 - \mathbb{E}[x_i^2]}{\mathrm{Var}(x_i^2)} \quad \text{and} \quad R_{i,j} = \frac{x_i x_j}{\mathrm{Var}(x_i)\,\mathrm{Var}(x_j)} - \frac{(x_i x_j - \mathbb{E}[x_i x_j])}{\mathrm{Var}(x_i x_j)} \ .$$

For complex designs where closed forms of these quantities are not readily available, these coefficients may be estimated to arbitrary precision via a Monte Carlo procedure (Fattorini, 2006). The following theorem shows that the proposed variance estimator is unbiased.

Theorem 4.10 (Unbiased Variance Estimator)**.** *Under the conditions in Lemmas 4.8 and 4.9, the variance estimator of the ERL point estimator is unbiased, i.e.* $\mathbb{E}[\widehat{\mathrm{Var}}(\hat{\tau})] = \mathrm{Var}(\hat{\tau})$.

When the conditions of Lemmas 4.8 and 4.9 do not hold, our proposed variance estimator will be ill-defined or biased. Indeed, it is possible that no unbiased variance estimator exists in such settings (Imbens and Rubin, 2015). In this case, one can replace the problematic $\mathrm{Var}(\hat{\tau}_i)$ and $\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)$ terms, which cannot be estimated directly, with upper bounds that can be estimated. Specifically, if the exposure x_i takes only two values, so Lemma 4.8 does not hold, then one can replace the variance of the individual treatment effect estimate with the second raw moment: $\mathrm{Var}(\hat{\tau}_i) = \mathbb{E}[\hat{\tau}_i^2] - \mathbb{E}[\hat{\tau}_i]^2 \leq \mathbb{E}[\hat{\tau}_i^2]$, as Aronow and Samii (2013) do when they invoke Young's inequality. We may thus replace the $y_i(\boldsymbol{z})Q_i$ terms in our variance estimator with $\hat{\tau}_i^2$. Similarly, if $\Psi_{i,j} = 0$ for some pair of outcome units $i, j \in V_o$ and Lemma 4.9 does not hold, then one can replace the corresponding covariance term with an upper bound obtained from Cauchy-Schwarz and AM–GM inequalities:

$$\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) \leq \sqrt{\mathrm{Var}(\hat{\tau}_i)\,\mathrm{Var}(\hat{\tau}_j)} \leq \frac{1}{2}\Big(\mathrm{Var}(\hat{\tau}_i) + \mathrm{Var}(\hat{\tau}_j)\Big) \ .$$

Under Assumptions 4.3 and 4.4, replacing one of these individual terms in this way leads to a positive bias of the normalized variance estimator on the order $\mathcal{O}(1/n)$. Thus, the variance estimator remains asymptotically unbiased as long as we apply these upper bounds to $o(n)$ terms. Even when the terms can be estimated without bias, it could still be preferable to use the bound here if the denominators of Q_i or $R_{i,j}$ are small, because small denominators will increase the variance of the estimator.

We may now use our variance estimator together with the asymptotic normality to construct well-motivated confidence intervals. We may estimate $1 - \alpha$ confidence

intervals by

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

where $\Phi^{-1} : [0,1] \to \mathbb{R}$ is the quantile function of the standard normal deviate. It is possible that the proposed variance estimator may take negative values; in particular, this may happen when the true variance is extremely small (on a normalized scale) or the sample size is small. When the variance estimator takes a negative value, this construction of confidence intervals is not well-defined. We suggest two possible alternatives here: the experimenter may either use the absolute value of the variance estimator under the square root, or the experimenter may use a more conservative variance estimate.

4.5 Analyzing ERL Without the Linear Response Assumption

Our previous analysis of the ERL estimator heavily relied on the linear response assumption. In this section, we show that without the linear response assumption, the ERL estimator may be interpreted as estimating a simple linear regression of the outcomes onto the exposures. The following theorem derives the expectation of the ERL estimator without the linear response assumption.

Theorem 4.11 (Arbitrary Responses)**.** *Assume that the potential functions are an arbitrary function of the exposures:* $y_i(\boldsymbol{z}) = y_i(x_i)$*. Then, the expectation of the* ERL *estimator is*

$$\mathbb{E}[\widehat{\tau}] = \frac{2}{n}\sum_{i=1}^{n} \hat{\beta}_i \ ,$$

where $\hat{\beta}_i$ *is the coefficient of the exposure* x_i *in an OLS regression of* y_i *on* x_i*:* $\hat{\beta}_i = \left(\frac{\mathrm{Cov}(x_i, y_i(x_i))}{\mathrm{Var}(x_i)}\right)$.

Theorem 4.11 shows that under a general (non-linear) response assumption, the ERL estimator may be interpreted as estimating the average of the slopes of the best linear fit of the outcome to the exposure. We emphasize that this regression cannot be run by the experimenter because the outcomes are not known. Nonetheless, Theorem 4.11 suggests that the ERL estimator may be interpreted, more generally, as estimating this regression-based estimand.

This result is related to several previous results within and outside causal inference. Realizing that most conditional expectation functions are not linear, statisticians and econometricians have advocated for an interpretation of linear regression as capturing an interpretable approximation of the underlying relationship between the outcome and the regressors (Chamberlain, 1984; Manski, 1991; Goldberger, 1991). Specifically for causal inference, Angrist (1998) highlights that when linear regression is used to estimate treatment effects in an observational setting, the estimator captures a

variance-weighted average of unit-level causal effects (see also Aronow and Samii, 2015 and Sloczynski, 2020). In a similar vein to these results, Theorem 4.11 shows that the ERL estimator captures a policy-relevant causal quantity even if the linear response assumption does not hold. The difference is that the effect it captures is a raw average over the units, and the approximation is with respect to each unit's response function.

Under the linear response assumption, this regression-based estimand is equal to the average total treatment effect (ATTE) defined in Section 4.2.3. However, these two estimands will not coincide for arbitrary response functions and designs. Aside from the linear response assumption, there are several scenarios where we will expect the ATTE and the regression-based estimand to be similar. The first scenario is when the design very closely approximates the Bernoulli design so that exposures have mean zero and concentrate around ± 1. When the design is exactly Bernoulli, one can verify that the regression-based estimand is exactly equal to the ATTE, which matches the intuition from the no-interference setting. The second scenario is when the response function is well-approximated by a linear function. An extensive investigation into formal conditions under which the regression-based estimand and the causal estimand (ATTE) are equivalent is beyond the scope of this work.

4.6 A Cluster Design for Targeting Exposure Distribution

In this section, we describe Exposure-Design, an independent cluster design which aims to improve precision of the ERL estimator by constructing a desirable exposure distribution. To this end, we first show in Section 4.6.1 that increasing the variance of exposures and decreasing the covariance between exposures can lead to improved precision of the ERL estimator in settings of interest. In Section 4.6.2, we present a clustering objective which aims to achieve such exposure distributions, thereby improving the precision of ERL estimator. Finally, we present a heuristic algorithm for optimizing this clustering objective in Section 4.6.3.

4.6.1 An ideal exposure distribution

Like all re-weighted linear estimators, the ERL estimator will incur a large mean squared error when the re-weighting terms are large. In particular, if the variance of an exposure $\mathrm{Var}(x_i(\boldsymbol{z}))$ is close to zero, the corresponding term of the estimator in (4.1) will become large, yielding a high mean squared error even though the estimator is unbiased. In general, experimenters should use designs for which the corresponding exposure variances are large.

However, large exposure variances should not be the only property of the exposure distribution that experimenters focus on. Consider a naive design that places equal probability on two treatment vectors: either all diversion units receive treatment

($\boldsymbol{z} = \mathbf{1}$) or all diversion units receive control ($\boldsymbol{z} = -\mathbf{1}$). Under this design, all of the exposure variances are 1, which is the largest possible variance. However, we observe either all of the treatment outcomes or all of the control outcomes, but never a mix of the two; in fact, the estimator itself takes only two values. Thus, the ERL estimator will suffer very large MSE under this design, despite the individual exposure variances being as large as possible. This raises the question: how should we construct a design that improve the precision of the ERL estimator?

This is a challenging task, since the precision of the ERL estimator depends on the unobserved outcomes. Indeed, a universally optimal design does not exist (Harshaw et al., 2021). However, we argue that a good heuristic is to construct the design so that the variance of the exposures are large and the covariance between most pairs of exposures are close to zero. As discussed at the end of Section 4.3.1, a design which directly targets these aspects of the exposure distribution may hope to ensure high precision of the ERL estimator under weaker growth conditions on the bipartite graph than those presented in our analysis.

To further motivate this heuristic, consider the scenario where all of the individual treatment effects are zero, i.e. the response functions are of the form $y_i(x_i) = \alpha_i$. Studies of these sort are sometimes called uniformity trails or A/A tests. In this scenario, one may derive the MSE of the ERL estimator as

$$\mathbb{E}[(\hat{\tau} - \tau)^2] = \frac{4}{n^2}\left[\sum_{i=1}^{m} \alpha_i^2 \frac{1}{\mathrm{Var}(x_i)} + 2\sum_{i<j} \alpha_i \alpha_j \frac{\mathrm{Cov}(x_i, x_j)}{\mathrm{Var}(x_i)\,\mathrm{Var}(x_j)}\right] .$$

As the individual variance terms increase, the first sum decreases. The effect of the second term depends on the sign of the product of intercepts, $\alpha_i \alpha_j$. Generally speaking, these intercepts are unknown to the experimenter. For the sake of this discussion, consider when the outcomes $y_i(\boldsymbol{z})$ are non-negative, in which case all intercepts α_i and their products $\alpha_i \alpha_j$ also are non-negative. In this case, decreasing the correlation between exposures would decrease the second term, leading to an overall decrease in the MSE of the ERL estimator.

4.6.2 Clustering objective for targeting exposure distribution

In the previous section, we argued for constructing a design so that the variance of exposures is large and the covariance between most exposures is small. However, as argued in Section 4.2.2, constructing a treatment distribution which realizes a desired exposure distribution is generally not possible due to overlapping structures in the bipartite graph. In this section, we present an optimization formulation for an independent cluster design which aims to achieve large exposure variance and small correlations between exposures, to the extent that this is possible given the bipartite graph.

We propose choosing a cluster design which maximizes the following objective

function:

$$\max_{\text{clustering } \mathcal{C}} \quad \sum_{i=1}^{n} \Bigg[\text{Var}(x_i) - \phi \sum_{i \neq j} \text{Cov}(x_i, x_j) \Bigg] \qquad \text{(Exposure-Design)}$$

The variance and covariance of the exposures above are with respect to the random treatment assignments of the corresponding independent cluster design. The first term in the objective is the sum of the exposure variances, so maximizing this term will encourage large exposure variances. The second term penalizes positive correlation between exposures, and maximizing it encourages small correlation. The correlation penalizing parameter $\phi \geq 0$ controls the relative emphasis between large exposure variances and small exposure correlations. When $\phi = 0$, then the emphasis is placed entirely on increasing individual exposure variance; this is typically undesirable, as the optimal solution is often a single cluster containing all diversion units, which results in the "naive" design where either all diversion units receive treatment or all diversion units receive control. Increasing ϕ places more emphasis on decorrelating exposures.

A key insight to solving the Exposure-Design formulation is that it may be reformulated as a *correlation clustering* problem, which is well-studied in the algorithms literature (Bansal et al., 2002; Swamy, 2004; Charikar et al., 2005). The existing computational understanding of these correlation clustering problems is another reason to use the Exposure-Design objective. The following proposition states the re-formulation of the Exposure-Design objective into the correlation clustering variant, denoted Corr-Clust.

Proposition 4.12. *For each pair of diversion units $i, j \in V_d$, define the value $\omega_{i,j} \in \mathbb{R}$ as*

$$\omega_{i,j} = (1+\phi) \sum_{k=1}^{m} w_{k,i} w_{k,j} - \phi \Big(\sum_{k=1}^{m} w_{k,i} \Big) \Big(\sum_{k=1}^{m} w_{k,j} \Big) \ , \qquad (4.2)$$

where $w_{k,i}$ is the weight of the edge between the kth outcome unit and the ith diversion unit. Exposure-Design *is equivalent to the following clustering problem:*

$$\max_{\textit{clusterings } \mathcal{C}} \quad \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} \ . \qquad \text{(Corr-Clust)}$$

Although Corr-Clust is a variant of the weighted maximization-type correlation clustering problems previously studied in the literature (Charikar et al., 2005; Swamy, 2004), is not equivalent to previously studied formulations in an approximation-preserving sense, as it takes positive and negative values. Given that weighted maximization correlation clustering is NP-Hard (Charikar et al., 2005), it is reasonable to presume that our formulation Corr-Clust is also computationally hard. However, these computational complexity considerations are beyond the scope of this work.

We remark that Exposure-Design places no explicit constraint on the number of clusters produced by the clustering algorithm. However, our analysis in Section 4.3.1

suggests that limiting the cluster sizes, and thereby reducing correlation between exposures, helps to achieve consistency and normality of the ERL estimator. This desirable cluster structure is not captured by the optimization problem itself, but we handle it through our local search heuristic described in Section 4.6.3. We also remark that the EXPOSURE-DESIGN objective does not directly minimize the MSE of the ERL estimator, but should instead be understood as a useful heuristic. Finally, we remark that it is impossible to induce a negative correlation between exposures in the class of independent cluster designs, so the second term of the objective attains its maximum when the exposures are uncorrelated.

The EXPOSURE-DESIGN is conceptually similar to the correlation-clustering based design presented in Pouget-Abadie et al. (2019), but it differs in several key ways. EXPOSURE-DESIGN provides experimenters the flexibility to trade-off larger exposure variances with more de-correlated exposures by setting the parameter ϕ. In contrast, the cluster design of Pouget-Abadie et al. (2019) focuses solely on the exposure variance by maximizing which is referred to as "empirical dose variance" in their paper. As we demonstrate in Appendix C.3, their objective is equal to ours when the trade-off parameter is set to $\phi = 1/(n-1)$. In this sense, their cluster design can be viewed as a specific instance of the more general EXPOSURE-DESIGN, where a greater emphasis is placed on maximizing the exposure variances. More importantly, the EXPOSURE-DESIGN presented in this chapter is designed to increase the precision of the ERL estimator, while the correlation-clustering based design of Pouget-Abadie et al. (2019) is motivated by the intuition that extreme exposures are helpful in this setting.

4.6.3 Local search heuristic for EXPOSURE-DESIGN

We now describe a local search heuristic for optimizing EXPOSURE-DESIGN. The local search is initialized with the singleton clustering and iteratively seeks to improve the clustering. In each iteration, the algorithm loops through random pairs of diversion units $i, j \in V_d$ and moves diversion unit j to the cluster currently containing diversion unit i if that change improves the objective value, subject to a user-defined constraint on the clusters. The local search algorithm is presented more formally below as

Algorithm 4.

Algorithm 4: Local Search($\boldsymbol{W}, \phi, k, T$, cluster constraints)

```
Initialize singleton clustering C = {{1}, {2}, ... {m}}
for iterations t = 1 ... T do
    Choose a uniformly random permutation π on the diversion units.
    for diversion units i ∈ π do
        Randomly select a diversion unit j with probability proportional to
         (W^⊤ W)_{i,j}
        Let C and C' be the clusters containing diversion units i and j,
         respectively.
        if moving j from cluster C' to C increases objective value & satisfies
         user-defined constraints then
            Move diversion unit j from cluster C' to C.
return clustering C
```

Given a diversion unit i, we use *wedge sampling* to randomly select another diversion unit j. that is, sampling j proportional to (Cohen and Lewis, 1999). We use the wedge sampling procedure because it encourages picking pairs of units for which $(\boldsymbol{W}^\intercal \boldsymbol{W})_{i,j} = \sum_{k=1}^{n} w_{k,i} w_{k,j}$ is large, which often results in a large correlation clustering weight $\omega_{i,j}$. Performance improvements are obtained by computing the correlation clustering weights $\omega_{i,j}$ only when they are needed to evaluate changes in the objective. In particular, the first term of (4.2) is an inner product whose computation scales with the sparsity of the bipartite graph and the second term is the product of sums which may be pre-computed.

Diversion unit j is moved into the cluster containing diversion unit i if two conditions are met: the objective increase and the user-defined cluster constraints are satisfied. We recommend that experimenters choose constraints which limit the cluster sizes in some way. For example, the experimenter may choose to constraint the number of diversion units within a cluster. In our implementation, we constrain the sum of the (unweighted) degrees of diversion units within a cluster to be a fixed fraction of the total number of edges. In this way, no cluster has too many outgoing edges to outcome units. Constraining the clusters in this way implicitly limits the amount of dependence between exposures, which is one of the key aspects underlying the design conditions in Assumption 4.4 of our analysis.

This local search algorithm is different from the one presented in Pouget-Abadie et al. (2019), which approximates the Gram matrix $\boldsymbol{W}^\intercal \boldsymbol{W}$ offline as the sum of a sparse matrix and a rank-one matrix, so that the algorithm works with an approximation to the objective. In contrast, our algorithm accepts and rejects changes based on the exact value of the objective. Relative to Elsner and Schudy (2009), this local search does not consider moving units to new empty clusters, nor does it consider merging clusters. Moves of the first type seem consistently unprofitable in our setting. As for merges, we find that the algorithm is able to essentially perform them

by moving one diversion unit at a time.

4.7 An Application to Online Marketplace Experiments

In this section, we apply our proposed methodology to a simulated marketplace experiment based on a product review dataset from the Amazon marketplace (McAuley et al., 2015; He and McAuley, 2016). The Amazon product review dataset contains 83 million reviews made by 121 thousand customers on 9.8 million items. In this application, we imagine running an experiment where we change the pricing mechanism of items in the marketplace, and are interested in how a customer's reported satisfaction is affected by this change in pricing mechanism. The items sold in the marketplace are the diversion units and the customers in the marketplace are the outcome units. An edge is present in the bipartite graph if a customer reviewed an item and all edges incident to an outcome unit are uniformly weighted. Thus, a customer's exposure is the unweighted average of the treatment status of the items they have previously reviewed.

In our simulated marketplace experiment, we generate potential outcomes via an exposure-response function. The outcomes themselves are the satisfaction score of a customer given their exposure; the responses in this study are simulated, but we can imagine that they are either reported directly by a customer or constructed based upon text analysis of the customer's review. In the case of a linear response, a positive slope indicates an increase in customer satisfaction as a result of the new pricing mechanism, while a negative slope indicates a decrease in satisfaction as a result of the new pricing mechanism.

We preprocess the Amazon produce review dataset for computational tractability in the same manner as Pouget-Abadie et al. (2019). We begin by removing customers that have reviewed fewer than 100 items. Next, we execute a balanced partitioning algorithm (Aydin et al., 2019) on the entire bipartite graph to create groups of customers and groups of items. After this preprocessing, we define the diversion units to be the item groups and the outcome units to be the customer groups. The resulting bipartite graph has 1 thousand outcome units, 2.4 million diversion units, and 7.1 million edges. We emphasize that this bipartite graph does not satisfy the growth conditions (specified in Section 4.3.1) required for consistency and normality under the broad class of designs captured by Assumption 4.4. In this sense, this application may be considered a test of the efficacy of the proposed Exposure-Design under weaker growth conditions.

We investigate the statistical properties of the ERL estimator, the variance estimator, and the resulting confidence intervals under various treatment designs in this application. In particular, we compare our proposed Exposure-Design to several existing designs: the Bernoulli design, the correlation clustering design of Pouget-

Abadie et al. (2019), and the balanced partitioning cluster design of Eckles et al. (2016a), as implemented by Aydin et al. (2019). Although the balanced partitioning design was not developed for the bipartite setting, we may expect it to achieve high precision estimates if the clustering produces decorrelated exposures with large variances.

We generate the potential outcomes in three simulations, where we vary the response functions that are used. The first two simulations feature linear response functions and the third simulation features a non-linear response function. We remark that although the parameters of the response functions are randomly chosen in our simulations, this random parameter draw is made only once and the outcomes themselves are fixed across all sampled assignments of all designs. These simulations are listed below.

- **(Mostly) Positive Treatment Effect.** In this simulation, we set almost all of the individual treatment effects to be positive across units, while varying the responses amongst the units. More precisely, we sample the slope terms as $\beta_i \sim \mathcal{N}(1, 1/4)$ and the intercept terms as $\alpha_i \sim \mathcal{N}(0, 1/8)$.

- **(Nearly) Zero Treatment Effect.** In this simulation, we set all the individual treatment effects close to zero, while varying the baseline outcomes. The outcomes are chosen to be mostly positive. More precisely, we sample the slope terms as $\beta_i \sim \mathcal{N}(0, 1/8)$ and the intercept terms as $\alpha_i \sim \mathcal{N}(2, 1/4)$.

- **Non-Linear Response.** In this simulation, we use a non-linear response function to specify the potential outcomes. In particular, the response of outcome unit i is $y_i(x_i) = 1 - x^2 + \alpha_i$, where $\alpha_i \sim \mathcal{N}(0, 1/8)$. Under this response, all individual treatment effects are 0. Because the linear response assumption is not satisfied, we should not expect our statistical analysis (unbiasedness, consistency, normality, etc) to hold exactly.

When using the Exposure-Design, we set the correlation penalty parameter to $\phi = 0.223$, chosen from a grid of 10 points between $[0, 2]$. The clustering itself is obtained using our local search heuristic presented in Section 4.6.3. Recall that the correlation clustering objective of Pouget-Abadie et al. (2019) may be obtained by setting $\phi = 1/(n-1)$. For this reason, we compute the corresponding cluster by running our local search heuristic with $\phi = 0.001 \approx 1/(n-1)$.

A summary of the main results from these simulations appears in Table 4.1. For each treatment design and simulation, we sample $20,000$ exposure vectors, compute the observed outcomes, and construct the corresponding ERL and variance estimators. Given the ERL and variance estimators, we construct the confidence intervals as described in Section 4.4, with absolute value corrections when the variance estimator takes a negative value. For each simulation and treatment design, we report the root mean square error (RMSE) of the ERL estimator, the average width of the 95% confidence intervals, and the coverage of the 95% confidence intervals.

		Exposure Design	Correlation Clustering	Balanced Partitioning	Bernoulli
Simulation 1	RMSE	0.049	0.088	0.073	0.659
	CI Width	0.219	0.328	0.283	2.576
	CI Coverage	91.8%	94.1%	93.7%	95.1%
Simulation 2	RMSE	1.81	2.05	1.85	43.83
	CI Width	7.03	8.00	7.21	190.8
	CI Coverage	94.7%	95.0%	94.6%	95.1%
Simulation 3	RMSE	0.86	0.90	0.78	24.37
	CI Width	3.47	3.82	3.39	95.15
	CI Coverage	95.6%	96.9%	96.9%	95.1%

Table 4.1: Simulation results

We draw particular attention to a few trends in these results. Exposure-Design achieves the smallest RMSE in the simulations which satisfy the linear response assumption. All cluster-based designs achieve significantly smaller RMSE than the Bernoulli design, which emphasizes the importance of carefully considering the exposure distribution when the growth conditions (specified in Section 4.3.1) are not satisfied. The confidence intervals in Simulation 1 under Exposure-Design cover below the nominal 95% level, indicating that either the sampling distribution of the ERL estimator isn't sufficiently approximated by a normal or the variance estimator isn't sufficiently concentrated. The confidence intervals in Simulation 3 cover slightly above the nominal 95% level, which is a result of conservative bias in the variance estimate due to non-linearity of the response.

Figure 4.1 contains histograms of the ERL estimator for each simulation and design, where the rows correspond to the designs and the columns correspond to the simulations. The dotted vertical line in the plot is the true ATTE. In all simulations, the distribution of the ERL estimator appears unimodal, centered around the ATTE, and (roughly) normal, which is empirical evidence that the normal approximation used to derive confidence intervals may be well-motivated for Exposure-Design and other cluster-based designs. This is to be expected for Simulations 1 and 2 where the linear response assumption holds, but is perhaps surprising in Simulation 3, which features a highly non-linear response. This unbiasedness may be explained by Theorem 4.11 in the following way: the quadratic responses in Simulation 3 yield zero treatment effect for all units. Although the best linear approximation to each quadratic response doe not well-approximate the quadratic response itself, the linear approximation has zero slope and so, in this sense, captures the ITE exactly.

Figure 4.2 contains histograms of the variance estimator for each simulation and treatment design, where the rows correspond to the designs and the columns correspond to the simulations. The dotted vertical line in the plot is the empirical estimate of the variance of the ERL estimator, computed from samples. The variance estimator is unbiased in Simulations 1 and 2, which aligns with Theorem 4.10; however, because

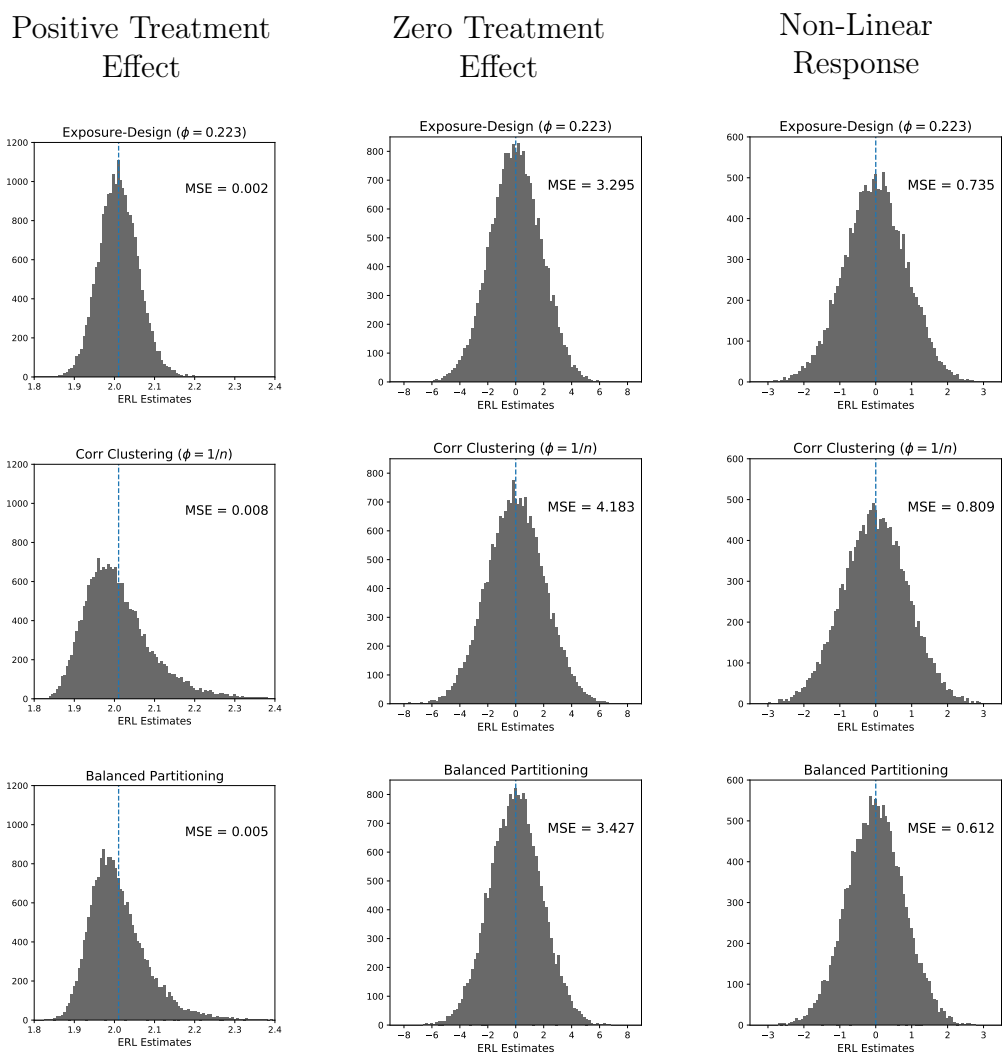

Figure 4.1: Histograms of the ERL estimator in simulations.

the empirical variance estimate is used, the blue line is close to (but not exactly) the true variance. Increasing the number of sampled exposure vectors decreases this error, but drawing more than 20 thousand samples is prohibitively expensive given the size of the data. In Simulation 1, the mean squared error of all cluster-based designs is so small that the variance estimator takes negative values. In Simulation 3, the response is highly non-linear so that the variance estimator incurs a positive bias, which results in a coverage slightly above the nominal level. Interestingly, the variance estimator under the balanced partitioning design is more concentrated around its mean, which is worth further rigorous investigation.

Figure 4.3 contains a plot for each simulation, where the mean squared error is plotted against the correlation penalizing parameter ϕ. The mean squared error of the balanced partitioning design appears as a dotted blue line. In Simulations 1 and 2 where the linear response assumption holds, there is a range of values of ϕ where

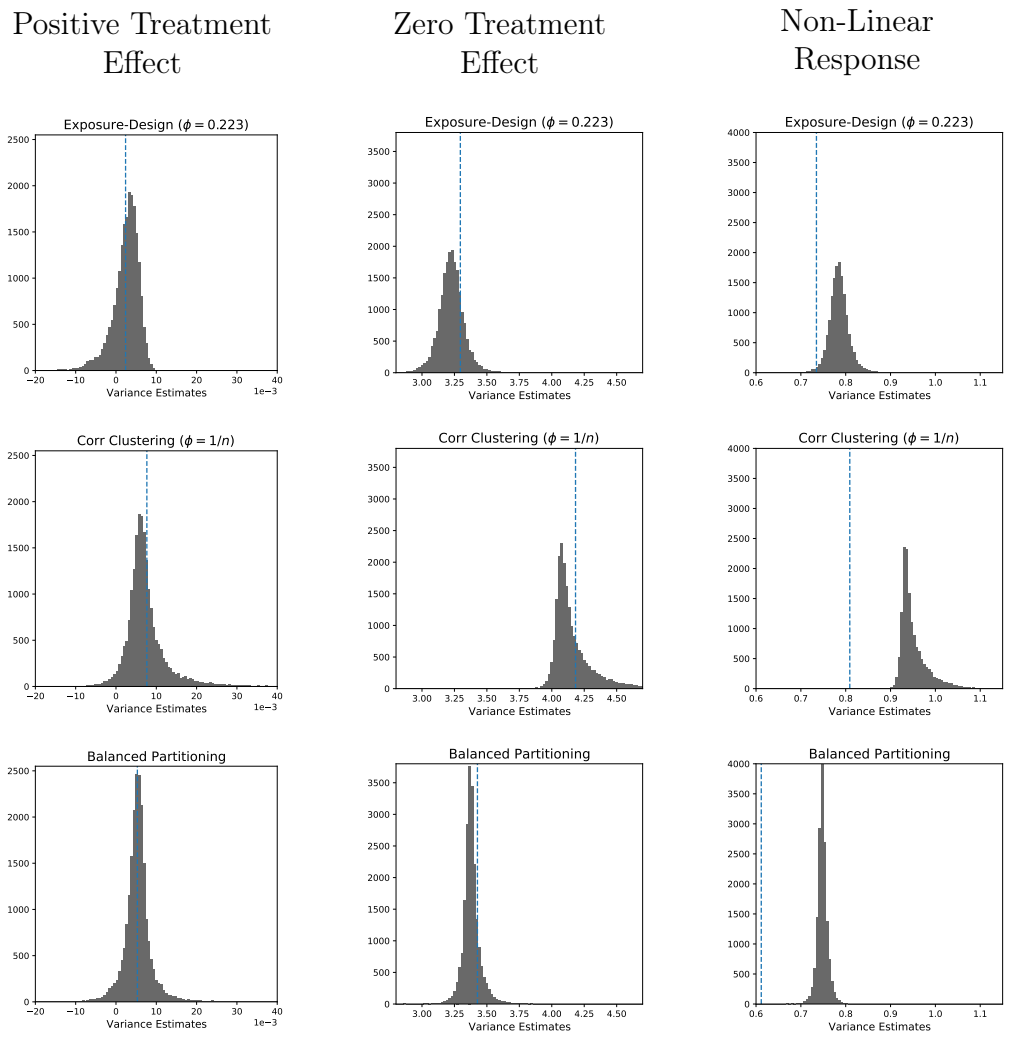

Figure 4.2: Histograms of the variance estimator in simulations.

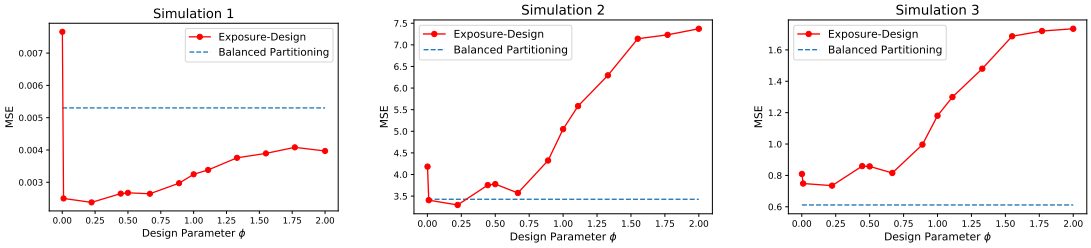

Figure 4.3: MSE of the ERL estimator as trade-off parameter ϕ is varied. First values of ϕ are .001 and .01.

Exposure-Design achieves lower mean squared error than the balanced partitioning design. In our simulations, the choice of $\phi \approx 1/4$ typically achieves lowest mean squared error. However, no design is optimal across all types of potential outcomes

(Harshaw et al., 2021) and so we encourage experimenters to select ϕ (and more generally, select designs) by running tests on simulated data.

4.8 Conclusion and Open Problems

In this chapter, we have presented methodological contributions towards estimating treatment effects in the linear-exposure response model for bipartite experiments: namely, the Exposure Reweighted Linear (ERL) estimator for consistent and unbiased estimation of average total treatment effect, an unbiased variance estimator which facilitates construction of normal-based confidence intervals, and the Exposure-Design that aims to increase precision of this estimator in various settings of interest.

When employing this design in practice, we recommend that the experimenter choose the value of the trade-off parameter ϕ by running simulations of the experiment using available models of the outcomes when possible. When this is not possible, we find in our simulations that setting $\phi \approx 1/4$ typically yields improvements in the precision of the ERL estimator over the previously correlation clustering design of Pouget-Abadie et al. (2019) where $\phi = 1/(n-1)$. We suspect that in most settings of interest, the ERL estimator will enjoy increased precision under any treatment design which (either explicitly or implicitly) ensures that exposures have large variance and are decorrelated.

There are several open questions suggested by this work. Answering any of the following methodological questions around the bipartite experimental framework will increase its relevance and applicability in practice.

- **Improved Designs**: *Construct a design such that consistency and asymptotic normality of the* ERL *estimator require weaker assumptions on the underlying bipartite graph.* Our current asymptotic analysis (consistency and asymptotic normality in Theorems 4.5 and 4.7, respectively) holds for a large class of designs, which requires stronger assumptions on the underlying bipartite graph. The proposed Exposure-Design achieves improved empirical performance for more dense bipartite graphs; however, it is heuristically motivated and there is room for improved designs with strong theoretical guarantees. A design which provably achieves even consistency under weaker conditions on the bipartite graph will likely result in a new distributional discrepancy formulation and sampling algorithm, thus constituting a major breakthrough.

- **Beyond Linear Response**: *Develop methods for valid inference in bipartite experiments which do not require the linear exposure-response assumption.* All methodology presented here relies upon the linear exposure-response assumption. A key question is to understand what sort of estimation and inference is possible under a general response function. For example, variance estimation under such general responses seems to require very different techniques

than those presented here, as they heavily rely on the parametric form of the response.

- **Misspecified Bipartite Graphs**: *Develop estimation and inference techniques that are robust to misspecification in the bipartite graph.* This work assumes that the bipartite graph is known to the experimenter. However, this assumption seems suspect in many applications where the bipartite graph may be constructed from historical data. Constructing designs and estimators which are (together) robust to minor misspecifications in the bipartite graph would greatly improve the applicability of the framework to practitioners. The results in Sävje (2021) regarding estimation of treatment effects under a misspecified exposure mapping might extend to this setting, but that remains to be shown.

Bibliography

Abadie, A. and Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, 91/92:175–187.

Angrist, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288.

Aral, S. and Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science*, 57(9):1623–1639.

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

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

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

Aronow, P. M. and Samii, C. (2015). Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60(1):250–267.

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

Aydin, K., Bateni, M., and Mirrokni, V. (2019). Distributed balanced partitioning via linear embedding. *Algorithms*, 12(8):162.

Bai, Y. (2019a). Optimality of matched-pair designs in randomized control trails. Available at SSRN: https://ssrn.com/abstract=3483834.

Bai, Y. (2019b). Randomization under permutation invariance. SSRN Preprint 3475147.

Baird, S., Bohren, J. A., McIntosh, C., and Özler, B. (2018). Optimal Design of Experiments in the Presence of Interference. *The Review of Economics and Statistics*, 100(5):844–860.

Banaszczyk, W. (1998). Balancing vectors and Gaussian measures of n-dimensional convex bodies. *Random Structures and Algorithms*, 12(4):351–360.

Bansal, N. (2014). Algorithmic aspects of combinatorial discrepancy. In Chen, W., Srivastav, A., and Travaglini, G., editors, *William ChenAnand SrivastavGiancarlo Travaglini*, chapter 6, pages 425–457. Springer.

Bansal, N., Blum, A., and Chawla, S. (2002). Correlation clustering. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, FOCS '02, page 238. IEEE Computer Society.

Bansal, N., Dadush, D., Garg, S., and Lovett, S. (2019). The Gram-Schmidt Walk: A cure for the Banaszczyk Blues. *Theory of Computing*, 15(21):1–27.

Basse, G., Ding, Y., and Toulis, P. (2019). Minimax crossover designs. ArXiv Preprint 1908.03531.

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

Basse, G. W., Soufiani, H. A., and Lambert, D. (2016). Randomization and the pernicious effects of limited budgets on auction experiments. In *Artificial Intelligence and Statistics*, pages 1412–1420. PMLR.

Beck, J. and Fiala, T. (1981). Integer-making theorems. *Discrete Applied Mathematics*, 3(1):1–8.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, Boston.

Bertsimas, D., Johnson, M., and Kallus, N. (2015). The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876.

Bhatia, R. (1997). *Matrix Analysis*. Springer, New York.

Blake, T. and Coey, D. (2014). Why marketplace experimentation is harder than it seems: The role of test-control interference. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 567–582.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, USA.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics*, 150(1):41–55.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Chamberlain, G. (1984). Panel data. In Griliches, Z. and Intriligator, M. D., editors, *Handbook of Econometrics*, volume 2, pages 1247–1318. Elsevier.

Charikar, M., Guruswami, V., and Wirth, A. (2005). Clustering with qualitative information. *J. Comput. Syst. Sci.*, 71(3):360–383.

Charikar, M., Newman, A., and Nikolov, A. (2011). Tight hardness results for minimizing discrepancy. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, page 1607–1614. Society for Industrial and Applied Mathematics.

Chazelle, B. (2000). *The Discrepancy Method: Randomness and Complexity*. Cambridge University Press, Cambridge.

Chin, A. (2019a). Central limit theorems via Stein's method for randomized experiments under interference. arXiv:1804.03105.

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

Cohen, E. and Lewis, D. D. (1999). Approximating matrix multiplication for pattern recognition tasks. *Journal of Algorithms*, 30(2):211 – 252.

Dadush, D., Garg, S., Lovett, S., and Nikolov, A. (2019). Towards a constructive version of banaszczyk's vector balancing theorem. *Theory of Computing*, 15(15):1–58.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21.

Doudchenko, N., Zhang, M., Drynkin, E., Airoldi, E., Mirrokni, V., and Pouget-Abadie, J. (2020). Causal inference with bipartite designs. *arXiv preprint arXiv:2010.02108*.

Drineas, P. and Mahoney, M. W. (2005). On the nyström method for approximating a gram matrix for improved kernel-based learning. *J. Mach. Learn. Res.*, 6:2153–2175.

Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: A toolkit. In Schultz, T. P. and Strauss, J. A., editors, *Handbook of Development Economics*, volume 4, chapter 61, pages 3895–3962. Elsevier.

Dunning, I., Huchette, J., and Lubin, M. (2017). Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320.

Eckles, D., Karrer, B., and Ugander, J. (2016a). Design and analysis of experiments in networks: Reducing bias from interference. *Journal of Causal Inference*, 5(1).

Eckles, D., Kizilcec, R. F., and Bakshy, E. (2016b). Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322.

Einav, L., Kuchler, T., Levin, J. D., and Sundaresan, N. (2011). Learning from seller experiments in online markets. Technical report, National Bureau of Economic Research.

Elsner, M. and Schudy, W. (2009). Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, page 19–27. Association for Computational Linguistics.

Fattorini, L. (2006). Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93(2):269–278.

Fisher, R. A. (1925). *Statistical Method for Research Workers*. Oliver & Boyd, Edinburgh.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513.

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

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

Fradkin, A. (2015). Search frictions and the design of online marketplaces. *Work. Pap., Mass. Inst. Technol.*

Fradkin, A. (2017). Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb. Available at SSRN: https://ssrn.com/abstract=2939084.

Freedman, D. A. (1975). On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118.

Gadbury, G. L. (2001). Randomization inference and bias of standard errors. *American Statistician*, 55(4):310–313.

Gittens, A. and Mahoney, M. (2013). Revisiting the nystrom method for improved large-scale machine learning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575, Atlanta, Georgia, USA. PMLR.

Goldberger, A. S. (1991). *A Course in Econometrics*. Harvard University Press, Cambridge.

Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics*, 5(2):263–275.

Gretton, A. (2020). Lecture notes in reproducing kernel hilbert spaces in machine learning. `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/rkhscourse.html`. Accessed: September 13, 2021.

Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., Cardin, N., Chandran, S., Chen, N., Coey, D., et al. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter*, 21(1):20–35.

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

Halloran, M. E. and Hudgens, M. G. (2016). Dependent happenings: A recent methodological review. *Current Epidemiology Reports*, 3(4):297–305.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, New York, second edition.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Higgins, M. J., Sävje, F., and Sekhon, J. S. (2016). Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences*, 113(27):7369–7376.

Higgins, M. J., Sävje, F., and Sekhon, J. S. (2015). Blocking estimators and inference under the neyman–rubin model.

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

Holtz, D., Lobel, R., Liskovich, I., and Aral, S. (2020). Reducing interference bias in online marketplace pricing experiments. *arXiv preprint arXiv:2004.12489*.

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

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

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, 27(24):4857–4873.

Imai, K., King, G., and Nall, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Statistical Science*, 24(1):29–53.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8580–8589, Red Hook, NY, USA. Curran Associates Inc.

Johari, R., Li, H., and Weintraub, G. (2020). Experimental design in two-sided platforms: An analysis of bias. In *Proceedings of the 21st ACM Conference on Economics and Computation*, page 851.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B*, 80(1):85–112.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(03):324–338.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967.

Kempton, R. (1997). Interference between plots. In *Statistical methods for plant variety evaluation*, pages 101–116. Springer.

Kondor, R. I. and Lafferty, J. D. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 315–322, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lesko, C. L., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., and Cole, S. R. (2017). Generalizing study results: A potential outcomes perspective. *Epidemiology*, 28:553–561.

Li, K.-C. (1983). Minimaxity for randomized designs: Some general results. *Annals of Statistics*, 11(1):225–239.

Li, X., Ding, P., and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162.

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

Liu, M., Mao, J., and Kang, K. (2020). Trustworthy online marketplace experimentation with budget-split design. *arXiv preprint arXiv:2012.08724*.

Lock Morgan, K. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282.

Manski, C. F. (1991). Regression. *Journal of Economic Literature*, 29(1):34–50.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies*, 60(3):531–542.

Marcus, A. W., Spielman, D. A., and Srivastava, N. (2015). Interlacing families ii: Mixed characteristic polynomials and the kadison–singer problem. *Annals of Mathematics*, 182:327–350.

Matoušek, J. (1999). *Geometric Discrepancy*. Springer, Berlin.

McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Menzel, K. and Imbens, G. (2021). A causal bootstrap. *Annals of Statistics*, in print.

Middleton, J. A. (2020). Unifying design-based inference: A new variance estimation principle. Mimeo.

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

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

Nutz, M. and Wang, R. (2020). The directional optimal transport.

Offer-Westort, M. and Dimmery, D. (2021). Experimentation for homogenous policy change. arXiv:2101.12318.

Ogburn, E. L., Sofrygin, O., Diaz, I., and van der Laan, M. J. (2020). Causal inference for social network data. arXiv:1705.08527.

Pashley, N. E. and Miratrix, L. W. (2019). Insights on variance estimation for blocked and matched pairs designs.

Pouget-Abadie, J., Aydin, K., Schudy, W., Brodersen, K., and Mirrokni, V. (2019). Variance reduction in bipartite experiments through correlation clustering. In *Advances in Neural Information Processing Systems*, pages 13309–13319.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Reiley, D. H. (2006). Field experiments on the effects of reserve prices in auctions: More magic on the internet. *The RAND Journal of Economics*, 37(1):195–211.

Robins, J. M. (1988). Confidence intervals for causal parameters. *Statistics in Medicine*, 7(7):773–785.

Rockafellar, R. T. (1993). Lagrange multipliers and optimality. *SIAM Review*, 35(2):183–238.

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

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688—701.

Rubin, D. B. (1980). Comment: Randomization analysis of experimental data. *Journal of the American Statistical Association*, 75(371):591.

Samii, C. and Aronow, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters*, 82(2):365–370.

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

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

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

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *In Proceedings of the Annual Conference on Computational Learning Theory*, pages 416–426.

Sinclair, B., McConnell, M., and Green, D. P. (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science*, 56(4):1055–1069.

Sloczynski, T. (2020). Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *Review of Economics and Statistics*, in press.

Smola, A. J. and Kondor, R. (2003). Kernels and regularization on graphs. In Schölkopf, B. and Warmuth, M. K., editors, *Learning Theory and Kernel Machines*, pages 144–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Spencer, J. (1985). Six standard deviations suffice. *Transactions of the American Mathematical Society*, 289(2):679–679.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer-Verlag, New York.

Stewart, G. W. (1998). *Matrix Algorithms: Vol. 1. Basic Decompositions*. Society for Industrial and Applied Mathematics, Philadelphia.

Strang, G. (2009). *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, fourth edition.

Struchiner, C. J., Halloran, M. E., Robins, J. M., and Spielman, A. (1990). The behaviour of common measures of association used to assess a vaccination programme under complex disease transmission patterns—a computer simulation study of malaria vaccines. *International journal of epidemiology*, 19(1):187–196.

Stuart, E. A., Cole, S. R., Bradshaw, C. R., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174:369–386.

Student (1923). On testing varieties of cereals. *Biometrika*, 15(3-4):271–293.

Student (1938). Comparison between balanced and random arrangements of field plots. *Biometrika*, 29(3/4):363–378.

Swamy, C. (2004). Correlation clustering: Maximizing agreements via semidefinite programming. SODA '04, page 526–527. Society for Industrial and Applied Mathematics.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38:239–266.

Toulis, P. and Kao, E. (2013). Estimation of causal peer influence effects. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 1489–1497.

Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia.

Udell, M., Mohan, K., Zeng, D., Hong, J., Diamond, S., and Boyd, S. (2014). Convex optimization in Julia. *SC14 Workshop on High Performance Technical Computing in Dynamic Languages*.

Vishwanathan, S., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *Journal of Machine Learning Research*, 11(40):1201–1242.

Wilk, M. B. (1955). The randomzation analysis of a generalized randomized block design. *Biometrika*, 42(1-2):70–79.

Wu, C.-F. (1981). On the robustness and efficiency of some randomized designs. *Annals of Statistics*, 9(6):1168–1177.

Yang, Y., Pilanci, M., and Wainwright, M. J. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991 – 1023.

Zigler, C. M. and Papadogeorgou, G. (2021). Bipartite causal inference with interference. *Statist. Sci.*, 36(1):109–123.

Contents

Appendix A

Appendix for Gram–Schmidt Walk Design

A.1 Analysis of the Gram–Schmidt Walk Algorithm

In this section, we restate the Gram–Schmidt Walk algorithm of Bansal et al. (2019) and present our analysis of the algorithm. We analyze the Gram–Schmidt Walk algorithm under more general conditions than what we consider in our analysis of the GSW-Design. At the end of the section, we discuss how the analysis of the Gram–Schmidt Walk algorithm extends to the GSW-Design.

We begin by restating the algorithm and introducing notation that will be used in the proofs. Next, we describe a formal connection to the Gram–Schmidt orthogonalization process which is also used in our proofs. We then provide proofs of the covariance bound (Theorem 2.12) and the subgaussian concentration (Theorem 2.19) of the Gram–Schmidt Walk algorithm. Finally, we discuss the extension of this analysis to the GSW-Design.

A.1.1 Gram–Schmidt Walk algorithm

In this section, we restate the Gram–Schmidt Walk algorithm using more detailed notation. This more detailed notation contains explicit references to the iteration index and will be used in the proofs in this supplement. Algorithm 5 below is the Gram–Schmidt Walk algorithm of Bansal et al. (2019). Randomizing the choice of pivots is not necessary for the algorithm or the analysis presented here, so we defer randomization of pivots to the discussion of the Gram–Schmidt Walk design in Section A.1.5. The algorithm presented in Section 4.6 sets the initial point $\boldsymbol{z}_1 = \boldsymbol{0}$.

We remark on some of the differences between the notation in Algorithm 5 here and the pseudo-code presented in the main body of Chapter 2. First, the Gram–Schmidt Walk algorithm takes as input arbitrary vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_n \in \mathbb{R}^m$. For purposes of analysis, we often assume that the ℓ_2 norms of these input vectors is at most 1. Second, in this version, which is identical to the algorithm developed

Algorithm 5: Gram–Schmidt Walk

Input : Vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_n \in \mathbb{R}^m$ arranged as columns in the matrix $\boldsymbol{B}$ and an initial point $\boldsymbol{z}_1 \in [-1, 1]^n$

Output: $\boldsymbol{z} \in \{\pm 1\}^n$

Set iteration index $t \leftarrow 1$ and alive set $\mathcal{A}_1 \leftarrow [n]$.

Set the first pivot $p_0 \leftarrow n$

while $\mathcal{A}_t \neq \emptyset$ **do**

  if $p_{t-1} \notin \mathcal{A}_t$ **then**

    Set the pivot p_t to the largest index in $\mathcal{A}_t$.

  else

    $p_t \leftarrow p_{t-1}$

  end

  Compute the step direction

$$\boldsymbol{u}_t \leftarrow \arg\min_{\boldsymbol{u} \in U} \|\boldsymbol{B}\boldsymbol{u}\|,$$

  where U is the set of all $\boldsymbol{u} \in \mathbb{R}^n$ such that $\boldsymbol{u}(p_t) = 1$ and $\boldsymbol{u}(i) = 0$ for all $i \notin \mathcal{A}_t$.

  Set $\delta_t^+ \leftarrow |\max \Delta|$ and $\delta_t^- \leftarrow |\min \Delta|$ where $\Delta = \{\delta \in \mathbb{R} : \boldsymbol{z}_t + \delta \boldsymbol{u}_t \in [-1, 1]^n\}$.

  Set the step size δ_t at random according to

$$\delta_t \leftarrow \begin{cases} \delta_t^+ & \text{with probability } \delta_t^- / (\delta_t^+ + \delta_t^-), \\ -\delta_t^- & \text{with probability } \delta_t^+ / (\delta_t^+ + \delta_t^-). \end{cases}$$

  Update the fractional assignment $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t + \delta_t \boldsymbol{u}_t$

  Update set of alive units $\mathcal{A}_{t+1} \leftarrow \{i \in [n] \mid |\boldsymbol{z}_t(i)| < 1\}$

  Increment the iteration index $t \leftarrow t + 1$

end

return $\boldsymbol{z} \leftarrow$ the final iterate $\boldsymbol{z}_{T+1}$

by Bansal et al. (2019), we do not choose the pivots at random. In fact, the only source of randomness in Algorithm 5 is the choice of step size δ_t at each iteration. In Section A.1.5, we demonstrate that selecting pivots uniformly at random from $\mathcal{A}_t$ is equivalent to randomly permuting the input order of the input vectors and running Algorithm 5. Finally, the notation presented here contains more reference to iteration indices. In particular, the notation of the pivot unit p_t, the alive set $\mathcal{A}_t$, and the choice of update steps δ_t^+, δ_t^- all feature the iteration index in the subscript. We also use the notation that $u_t(i)$ denotes the ith coordinate of the vector $\boldsymbol{u}$ at time t.

We denote the (random) number of iterations by T. We now introduce a notational convention which improves the clarity of some further analysis. Because the number of iterations T is always at most n by Lemma 2.10, we may suppose that the algorithm runs for exactly n iterations and that for iterations $t > T$, we set the update direction $\boldsymbol{u}_t = \mathbf{0}$ and the step size $\delta_t = 0$. The same vector $\boldsymbol{z}$ is returned and the output distribution of the algorithm is unchanged. We remark that this convention is used sparingly throughout the analysis and does not change the algorithm.

The concept of pivot phases was central to the analysis in Bansal et al. (2019) and it remains a central part of the analysis presented here as well. For each unit $i \in [n]$, we define the *pivot phase* S_i to be the set of iterations for which unit i is the pivot, i.e.

$$S_i = \{t : p_t = i\}.$$

During a particular run of the algorithm, the pivot phase S_i may be empty if unit i is not chosen as a pivot unit during that run.

During the course of the algorithm, a unit $i \in [n]$ is said to be *alive* if $|\boldsymbol{z}_t(i)| < 1$ and *frozen* otherwise. This is the convention is used by Bansal et al. (2019) and it reflects that fact that once a unit is frozen, its fractional assignment becomes integral and it is no longer updated. The set $\mathcal{A}_t$ is referred to as the *alive set* because it contains all alive units at the beginning of iteration t. We refer to the vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_n$ as the input vectors. We may slightly abuse our terminology and call an input vector $\boldsymbol{b}_i$ alive or frozen when we mean that the corresponding unit i is alive or frozen.

We say that a unit i is *decided by the algorithm* when it is either selected as the pivot (Lines 2 or 5) or frozen without being chosen as the pivot (Line 12). Throughout the proofs below, we often condition on the previous random decisions made by the algorithm. We use Δ_i to denote all the random decisions made by the algorithm up to and including when unit i was decided by the algorithm. There is, however, some care to be taken in this definition to distinguish between units which are chosen as pivots and those which are not. If i is chosen as a pivot at the beginning of iteration t, then Δ_i includes all previous choices of step sizes $\delta_1 \ldots \delta_{t-1}$. If i is frozen at the end of iteration t without being chosen as the pivot, then Δ_i includes all choices of step sizes $\delta_1 \ldots \delta_t$. Other types of conditioning will be presented throughout the proofs as the needs arise.

A.1.2 Connection to Gram–Schmidt orthogonalization

A key aspect in our analysis of the Gram–Schmdit Walk algorithm is a Gram–Schmidt orthogonalization applied to a random re-ordering of the input vectors. We use the randomized Gram–Schmidt orthogonalization to obtain the tight bounds on the covariance matrix and the subgaussian constant in Theorems 2.12 and 2.19, respectively. In this section, we describe this connection in detail, providing additional notation and several technical lemmas which will be used in the proofs of Theorems 2.12 and 2.19.

Before continuing, we make two remarks regarding the randomized Gram–Schmidt orthogonalization. First, we emphasize that this re-ordering and orthogonalization is only for the purposes of analysis and is not executed by the algorithm. We also remark that although Bansal et al. (2019) discuss how the Gram–Schmidt Walk algorithm was inspired by Gram–Schmidt orthogonalization, an explicit connection is not made in that paper. This is one of the technical differences in our analysis which allow us to obtain tighter bounds.

We begin this discussion by first describing the randomized re-ordering of the input vectors and then defining the Gram–Schmidt Orthogonalization processes applied to this re-ordering. Let us introduce the notation of the re-ordering. The inputs vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \dots \boldsymbol{b}_n \in \mathbb{R}^m$ will be re-ordered as

$$\boldsymbol{b}_{\boldsymbol{\sigma}(1)}, \boldsymbol{b}_{\boldsymbol{\sigma}(2)}, \dots \boldsymbol{b}_{\boldsymbol{\sigma}(n)} \ ,$$

where $\boldsymbol{\sigma}$ is a bijection mapping positions in the re-ordering to the units. Formally, $\boldsymbol{\sigma} : [n] \to [n]$ and to avoid confusion in this notation, we reserve the symbol r for a position in the re-ordering and the symbol i for a unit. In this way, we write $\boldsymbol{\sigma}(r) = i$ to mean that the rth position in the re-ordering is occupied by unit i. We may also refer to the position of a specific unit in the re-ordering using the inverse function $\boldsymbol{\sigma}^{-1}$. That is, $\boldsymbol{\sigma}^{-1}(i) = r$ means that the unit i is assigned to position r in the re-ordering.

The re-ordering we consider is random and it is defined by the random choices made in the algorithm. Recall that a unit i is decided by the algorithm when it is either selected as the pivot (Lines 2 or 5) or frozen without being chosen as the pivot (Line 12). The ordering of the units $\boldsymbol{\sigma}(1), \boldsymbol{\sigma}(2), \dots \boldsymbol{\sigma}(n)$ will be the *reverse order* in which they are decided, breaking ties arbitrarily. In this way, as the algorithm decides units at each iteration, the randomized re-ordering is determined in reverse order. For example, the first unit to be decided is the first pivot unit p_1 so that $\boldsymbol{\sigma}(n) = p_1 = n$. If a single unit $j \neq p_1$ is frozen in the first iteration, then this is the next unit decided by the algorithm, in which case it is second to last in the re-ordering, i.e. $\boldsymbol{\sigma}(n-1) = j$. On the other hand, if only the pivot p_1 is frozen in the first iteration, the next unit decided by the algorithm is the next pivot, which is p_2. In this case, $\boldsymbol{\sigma}(n-1) = p_2$.

Next, we introduce the Gram–Schmidt orthogonalization process on this randomized re-ordering of the input vectors. The Gram–Schmidt orthogonalization process is a method to construct a sequence of orthonormal vectors which form a basis for

the span of a given set of vectors. For our problem at hand, we denote this sequence of orthonormal basis vectors by

$$\boldsymbol{w}_{\boldsymbol{\sigma}(1)}, \boldsymbol{w}_{\boldsymbol{\sigma}(2)}, \ldots \boldsymbol{w}_{\boldsymbol{\sigma}(n)}.$$

They are recursively defined by the Gram–Schmidt orthogonalization process

$$\boldsymbol{w}_{\boldsymbol{\sigma}(1)} = \frac{\boldsymbol{b}_{\boldsymbol{\sigma}(1)}}{\|\boldsymbol{b}_{\boldsymbol{\sigma}(1)}\|} \quad \text{and} \quad \boldsymbol{w}_{\boldsymbol{\sigma}(r)} = \frac{\boldsymbol{b}_{\boldsymbol{\sigma}(r)} - \boldsymbol{A}_r \boldsymbol{b}_{\boldsymbol{\sigma}(r)}}{\left\|\boldsymbol{b}_{\boldsymbol{\sigma}(r)} - \boldsymbol{A}_r \boldsymbol{b}_{\boldsymbol{\sigma}(r)}\right\|} \quad \text{for } r = 2, \ldots n,$$

where $\boldsymbol{A}_r = \sum_{s<r} \boldsymbol{w}_{\boldsymbol{\sigma}(s)} \boldsymbol{w}_{\boldsymbol{\sigma}(s)}^{\intercal}$ is the projection onto the span of the first $r-1$ input vectors $\boldsymbol{b}_{\boldsymbol{\sigma}(1)} \ldots \boldsymbol{b}_{\boldsymbol{\sigma}(r-1)}$. Because the random re-ordering of the input vectors is determined by the random choices of $\delta_1 \ldots \delta_n$ in the algorithm, the random sequence $\boldsymbol{w}_{\boldsymbol{\sigma}(1)} \ldots \boldsymbol{w}_{\boldsymbol{\sigma}(n)}$ is also determined by the random choices made by the algorithm. Regardless of the randomization, this sequence of vectors forms an orthonormal basis for the span of the input vectors. Moreover, while the vector $\boldsymbol{w}_{\boldsymbol{\sigma}(r)}$ depends on the set of vectors $\{\boldsymbol{b}_{\boldsymbol{\sigma}(1)}, \ldots, \boldsymbol{b}_{\boldsymbol{\sigma}(r-1)}\}$, it does not depend on their order. For further reading on the Gram–Schmidt orthogonalization process, we refer readers to Chapter 4 of Strang (2009).

The main benefit of using this Gram–Schmidt orthogonalization process is that we can cleanly analyze the behavior of the algorithm within pivot phases. In particular, it provides a way to partition the span of the input vectors into orthogonal subspaces $V_1, V_2, \ldots V_n$ corresponding to each of the n units. These subspaces are defined by the algorithm's random choices within the corresponding unit's pivot phase. We begin by defining the subspaces for units that are chosen as pivots. Let i be a unit which is chosen as pivot and assume it has position $r = \boldsymbol{\sigma}^{-1}(i)$ in the reordering so that the $k+1$ vectors which are decided during this pivot phase appear in the ordering as $\boldsymbol{b}_{\boldsymbol{\sigma}(r-k)}, \boldsymbol{b}_{\boldsymbol{\sigma}(r-k+1)}, \ldots \boldsymbol{b}_{\boldsymbol{\sigma}(r)}$. The subspace $V_i \subset \mathbb{R}^m$ is defined to be the span of the vectors $\boldsymbol{b}_{\boldsymbol{\sigma}(r-k)}, \boldsymbol{b}_{\boldsymbol{\sigma}(r-k+1)}, \ldots \boldsymbol{b}_{\boldsymbol{\sigma}(r)}$ after they have been projected orthogonal to $\boldsymbol{b}_{\boldsymbol{\sigma}(1)}, \boldsymbol{b}_{\boldsymbol{\sigma}(2)}, \ldots, \boldsymbol{b}_{\boldsymbol{\sigma}(r-k-1)}$. As the set $\{\boldsymbol{\sigma}(1), \ldots, \boldsymbol{\sigma}(r-k-1)\}$ is determined at this time, the projection is well-defined. The vectors

$$\boldsymbol{w}_{\boldsymbol{\sigma}(r-k)}, \boldsymbol{w}_{\boldsymbol{\sigma}(r-k+1)}, \ldots, \boldsymbol{w}_{\boldsymbol{\sigma}(r)}$$

form an orthonormal basis for the subspace V_i and the projection matrix onto this subspace is

$$\boldsymbol{P}_i = \sum_{s=0}^{k} \boldsymbol{w}_{\boldsymbol{\sigma}(r-s)} \boldsymbol{w}_{\boldsymbol{\sigma}(r-s)}^{\intercal}.$$

If a unit i is never chosen as a pivot unit, then V_i is the zero subspace and so the projection matrix P_i is the zero matrix. We remark that these subspaces and projection matrices are the ones referenced in the proof sketches of Theorems 2.12 and 2.19.

The following lemma follows directly from the definition of the subspaces but may also be verified by orthonormality of the vector sequence produced by Gram–Schmidt

orthogonalization.

Lemma A.1. *The subspaces are $V_1, V_2, \dots V_n$ are orthogonal and their union is* $\text{span}\{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots \boldsymbol{b}_n\}$. *Equivalently, the corresponding projection matrices $\boldsymbol{P}_1 \dots \boldsymbol{P}_n$ satisfy*

$$\sum_{i=1}^{n} \boldsymbol{P}_i = \boldsymbol{P},$$

where $\boldsymbol{P}$ is the projection matrix onto $\text{span}\{\boldsymbol{b}_1, \boldsymbol{b}_2, \dots \boldsymbol{b}_n\}$.

Next, we will show that the fractional balance update $\boldsymbol{B}\boldsymbol{u}_t$ is contained in the subspace corresponding to the current pivot, V_{p_t}. We will show a stronger property, but in order to make these statements precise, we need additional notation which connects an iteration t with the re-ordered positions of the units that have already been decided during in the current pivot phase. We define ℓ_t and g_t to be the least and greatest re-ordering positions that were decided during the current pivot phase before Line 9 at iteration t. The first unit to be decided in any pivot phase is the pivot unit. Thus the greatest re-ordering position of any unit which was decided during the current pivot phase is $g_t = \boldsymbol{\sigma}^{-1}(p_t)$. Note that when we arrive at Line 9, $\mathcal{A}_t \setminus p_t$ is the set of units which have not yet been decided. Thus, these are the units which will appear earliest in the re-ordering (although their ordering is not yet determined) and so we have that $\ell_t = |\mathcal{A}_t \setminus p_t| + 1 = |\mathcal{A}_t|$. In the first iteration of a pivot phase, we have $\ell_t = g_t$ because only the pivot has been decided before Line 9 at this iteration.

Using this notation, at Line 9 of iteration t, the input vectors whose units have been decided during the current pivot phase are

$$\boldsymbol{b}_{\boldsymbol{\sigma}(\ell_t)}, \boldsymbol{b}_{\boldsymbol{\sigma}(\ell_t+1)}, \dots \boldsymbol{b}_{\boldsymbol{\sigma}(g_t)}.$$

The next lemma demonstrates that the fractional update $\boldsymbol{B}\boldsymbol{u}_t$ is the projection of the pivot onto the subspace spanned by $\boldsymbol{w}_{\boldsymbol{\sigma}(\ell_t)}, \boldsymbol{w}_{\boldsymbol{\sigma}(\ell_t+1)}, \dots \boldsymbol{w}_{\boldsymbol{\sigma}(g_t)}$.

Lemma A.2. *At each iteration t, we can write $\boldsymbol{B}\boldsymbol{u}_t$ in the orthonormal basis $\boldsymbol{w}_{\boldsymbol{\sigma}(1)} \dots \boldsymbol{w}_{\boldsymbol{\sigma}(n)}$ as*

$$\boldsymbol{B}\boldsymbol{u}_t = \sum_{r=\ell_t}^{g_t} \big\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t} \big\rangle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}.$$

Proof. Recall that the step direction $\boldsymbol{u}_t$ is determined by a least squares problem. That is, the undecided coordinates of the step direction, $u_t(\mathcal{A}_t \setminus p_t)$, are the minimizers of the least squares program

$$u_t(\mathcal{A}_t \setminus p_t) = \underset{u_i : i \in \mathcal{A}_t \setminus p_t}{\arg\min} \left\| \boldsymbol{b}_{p_t} + \sum_{i \in \mathcal{A}_t \setminus p_t} u_i \boldsymbol{b}_i \right\|^2.$$

Because the step direction is the minimizer, it must satisfy the normal equations

$$\boldsymbol{B}\boldsymbol{u}_t = \boldsymbol{b}_{p_t} - \boldsymbol{A}_t \boldsymbol{b}_{p_t},$$

where $\boldsymbol{A}_t$ is the projection matrix onto the span of the alive vectors which are not the pivot. That is, $\boldsymbol{b}_i$ for i in $\mathcal{A}_t \setminus p_t = \{\boldsymbol{\sigma}(1), \dots, \boldsymbol{\sigma}(\ell_t) - 1\}$. By the construction of the re-ordering and the Gram–Schmidt orthogonalization, we have that $\boldsymbol{A}_t = \sum_{s<\ell_t} \boldsymbol{w}_{\boldsymbol{\sigma}(s)} \boldsymbol{w}_{\boldsymbol{\sigma}(s)}^\intercal$. Writing the fractional balance update $\boldsymbol{B}\boldsymbol{u}_t$ in the orthonormal basis, we have that

$$\begin{aligned}
\boldsymbol{B}\boldsymbol{u}_t &= \sum_{r=1}^{n} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{B}\boldsymbol{u}_t \rangle \boldsymbol{w}_{\boldsymbol{\sigma}(r)} && \text{(orthonormal basis)} \\
&= \sum_{r=1}^{n} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t} - \boldsymbol{A}_t \boldsymbol{b}_{p_t} \rangle \boldsymbol{w}_{\boldsymbol{\sigma}(r)} && \text{(normal equations)} \\
&= \sum_{r=1}^{n} \Big[\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t} \rangle - \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{A}_t \boldsymbol{b}_{p_t} \rangle \Big] \boldsymbol{w}_{\boldsymbol{\sigma}(r)} && \text{(linearity)} \\
&= \sum_{r=1}^{n} \Big[\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t} \rangle - \langle \boldsymbol{A}_t \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t} \rangle \Big] \boldsymbol{w}_{\boldsymbol{\sigma}(r)}. && \text{(projection matrix, } \boldsymbol{A}_t^\intercal = \boldsymbol{A}_t)
\end{aligned}$$

We now examine each term in this sum. If $r < \ell_t$ then $\boldsymbol{A}_t \boldsymbol{w}_{\boldsymbol{\sigma}(r)} = \boldsymbol{w}_{\boldsymbol{\sigma}(r)}$ because $\boldsymbol{w}_{\boldsymbol{\sigma}(r)}$ is a vector in the subspace associated with the projection $\boldsymbol{A}_t$. Thus, the two terms in the bracket are the same, so the terms corresponding to $r < \ell_t$ are zero and do not contribute to the sum. If $r \geq \ell_t$, then by the construction of the re-ordering and Gram–Schmidt orthogonalization, $\boldsymbol{w}_{\boldsymbol{\sigma}(r)}$ is orthogonal to the subspace corresponding to $\boldsymbol{A}_t$ and so $\boldsymbol{A}_t \boldsymbol{w}_{\boldsymbol{\sigma}(r)} = 0$. This means that for $\ell_t \leq r \leq g_t$, the second term in the brackets is zero, and only the first term in brackets contributes to the sum. On the other hand, if $r > g_t$, then by the re-ordering and Gram–Schmidt orthogonalization, $\boldsymbol{w}_{\boldsymbol{\sigma}(r)}$ is orthogonal to $\boldsymbol{b}_{\boldsymbol{\sigma}(g_t)} = \boldsymbol{b}_{p_t}$. In this case, both terms in the brackets are zero and the terms corresponding to $r > g_t$ contribute nothing to the sum. Thus, we have shown that

$$\boldsymbol{B}\boldsymbol{u}_t = \sum_{r=\ell_t}^{g_t} \big\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t} \big\rangle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}. \qquad \square$$

A.1.3 Covariance bound (Theorem 2.12)

This section contains a proof of an extended version of the covariance bound in Theorem 2.12. We begin by deriving a form of the covariance matrix of the assignment vector in terms of the update quantities in the algorithm.

Lemma A.3. *The covariance matrix of the assignment vector is given by*

$$\operatorname{Cov}(\boldsymbol{z}) = \mathbb{E}\Big[\sum_{t=1}^{T} \delta_t^2 \boldsymbol{u}_t \boldsymbol{u}_t^\intercal \Big].$$

Proof. First, observe that

$$\mathrm{Cov}(\boldsymbol{z}) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal] - \mathbb{E}[\boldsymbol{z}]\,\mathbb{E}[\boldsymbol{z}]^\intercal = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal] - \boldsymbol{z}_1\boldsymbol{z}_1^\intercal$$

where the second equality uses $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{z}_1$, which in a consequence of the martingale property (Lemma 2.7). By the update rule $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t + \delta_t \boldsymbol{u}_t$,

$$\boldsymbol{z}_{t+1}\boldsymbol{z}_{t+1}^\intercal = (\boldsymbol{z}_t + \delta_t\boldsymbol{u}_t)(\boldsymbol{z}_t + \delta_t\boldsymbol{u}_t)^\intercal = \boldsymbol{z}_t\boldsymbol{z}_t^\intercal + \delta_t\big(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal\big) + \delta_t^2\boldsymbol{u}_t\boldsymbol{u}_t^\intercal.$$

Iteratively applying this over all iterations $t \in \{1, 2, \ldots\}$ and using that the returned vector is $\boldsymbol{z} = \boldsymbol{z}_{T+1}$, we have that

$$\boldsymbol{z}\boldsymbol{z}^\intercal = \boldsymbol{z}_{T+1}\boldsymbol{z}_{T+1}^\intercal = \boldsymbol{z}_1\boldsymbol{z}_1^\intercal + \sum_{t=1}^{T}\delta_t\big(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal\big) + \sum_{t=1}^{T}\delta_t^2\boldsymbol{u}_t\boldsymbol{u}_t^\intercal.$$

Substituting this expression of $\boldsymbol{z}\boldsymbol{z}^\intercal$ into $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal]$ in the earlier covariance calculation, we obtain that

$$\mathrm{Cov}(\boldsymbol{z}) = \mathbb{E}\left[\sum_{t=1}^{T}\delta_t^2\boldsymbol{u}_t\boldsymbol{u}_t^\intercal\right] + \mathbb{E}\left[\sum_{t=1}^{T}\delta_t\big(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal\big)\right] \tag{A.1}$$

We will now show that the last term is zero because the step size δ_t is zero in expectation. By linearity of expectation and using the convention that the algorithm runs for n iterations with $\delta_t = 0$ and $\boldsymbol{u}_t = \boldsymbol{0}$ for $t > T$,

$$\mathbb{E}\left[\sum_{t=1}^{T}\delta_t\big(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal\big)\right] = \sum_{t=1}^{n}\mathbb{E}\big[\delta_t\big(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal\big)\big]$$

For a fixed iteration t, consider the individual term $\mathbb{E}[\delta_t(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal)]$ in the sum above. Observe that if we condition on all previous random decisions made by the algorithm before step size δ_t is chosen (i.e. choices of step sizes $\delta_1 \ldots \delta_{t-1}$), then the step direction $\boldsymbol{u}_t$ and fractional assignment $\boldsymbol{z}_t$ are both determined, so that $\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal$ is a deterministic quantity. In this way, δ_t is conditionally independent of $\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal$ conditioned on all previous random decisions made by the algorithm. Using the fact that the expected step size δ_t is zero, we have that

$$\mathbb{E}[\delta_t\big(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal\big) \mid \delta_1 \ldots \delta_{t-1}] = \big(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal\big) \cdot \mathbb{E}[\delta_t \mid \delta_1 \ldots \delta_{t-1}] = 0$$

for all iterations t. By the law of total expectation, $\mathbb{E}[\delta_t(\boldsymbol{u}_t\boldsymbol{z}_t^\intercal + \boldsymbol{z}_t\boldsymbol{u}_t^\intercal)] = 0$ and so that the second term in (A.1) is zero. □

Next, we prove a lemma stating that the expected sum of the squared step sizes in the remainder of a pivot phase is not too large in expectation. To do this, we introduce notation that connects a position in the re-ordering to the subsequent iterations in a

pivot phase. For each position r in the re-ordering, we define

$$L_r = \{t : \ell_t \leq r \leq g_t\}.$$

The set L_r allows us to discuss what happens in the remaining iterations of a pivot phase after the unit in position r has been decided. For example, if a unit i is chosen as the pivot and assigned to position r, then L_r is the entire pivot phase S_i. If a non-pivot unit i is frozen and assigned to position r, then L_r are the remaining iterations in the pivot phase. Note that L_r may be empty if a non-pivot unit is frozen along with pivot at the last iteration of the pivot phase. We are now ready to state a lemma on the expected sum of the squared step sizes throughout the remainder of a pivot phase.

Lemma A.4. *For each $r \in [n]$, conditional on the random decisions made up until unit $\boldsymbol{\sigma}(r)$ is decided, the expected sum of squared step sizes in the remainder of its pivot phase is at most one. That is, for each unit $i \in [n]$ with re-ordering position $r = \boldsymbol{\sigma}^{-1}(i)$,*

$$\mathbb{E}\left[\sum_{t \in L_r} \delta_t^2 \,\middle|\, \Delta_{\boldsymbol{\sigma}(r)}\right] \leq 1.$$

Proof. Because only one pivot phase is being considered, we drop the iteration subscripts here and write the pivot as p. Recall that $\Delta_{\boldsymbol{\sigma}(r)}$ denotes all the random decisions made by the algorithm up to and including when unit i was decided by the algorithm. If L_r is empty, then the statement is trivially true. Otherwise, L_r is a (random) contiguous set of iterations $t_0, t_0 + 1, \ldots t_0 + k$, where $t_0 + k$ is the last iteration in the pivot phase. Because the pivot phase terminates when the pivot p is frozen, $|\boldsymbol{z}_{t_0+k}(p)| = 1$. It follows that

$$\begin{aligned}
1 - \boldsymbol{z}_{t_0}(p)^2 &= \boldsymbol{z}_{t_0+k}(p)^2 - \boldsymbol{z}_{t_0}(p)^2 && (|\boldsymbol{z}_{t_0+k}(p)| = 1) \\
&= \sum_{s=0}^{k-1} \left[\boldsymbol{z}_{t_0+s+1}(p)^2 - \boldsymbol{z}_{t_0+s}(p)^2\right] && \text{(telescoping sum)} \\
&= \sum_{s=0}^{k-1} \left[\left(\boldsymbol{z}_{t_0+s}(p) + \delta_{t_0+s}\boldsymbol{u}_{t_0+s}(p)\right)^2 - \boldsymbol{z}_{t_0+s}(p)^2\right] && \text{(update rule)} \\
&= \sum_{s=0}^{k-1} \left[\delta_{t_0+s}^2 \boldsymbol{u}_{t_0+s}(p)^2 + 2\delta_{t_0+s}\boldsymbol{u}_{t_0+s}(p)\boldsymbol{z}_{t_0+s}(p)\right] && \text{(cancelling terms)}
\end{aligned}$$

Taking conditional expectations of both sides and using linearity of expectation, we have that

$$1 - \boldsymbol{z}_{t_0}(p)^2 = \mathbb{E}\left[\sum_{t \in L_r} \delta_t^2 \,\middle|\, \Delta_{\boldsymbol{\sigma}(r)}\right] + 2\,\mathbb{E}\left[\sum_{t \in L_r} \delta_t \boldsymbol{u}_t(p)\boldsymbol{z}_t(p) \,\middle|\, \Delta_{\boldsymbol{\sigma}(r)}\right], \tag{A.2}$$

because the left hand side is a deterministic quantity under this conditioning. We now seek to show that the second term on the right hand side is zero. To this end, observe that we may extend the sum from iterations $t \in L_r$ to all remaining iterations because $\boldsymbol{u}_t(p) = 0$ for iterations t after the current pivot phase, i.e.,

$$\mathbb{E}\left[\sum_{t\in L_r} \delta_t \boldsymbol{u}_t(p)\boldsymbol{z}_t(p)\middle|\Delta_{\boldsymbol{\sigma}(r)}\right] = \mathbb{E}\left[\sum_{t\geq t_0} \delta_t \boldsymbol{u}_t(p)\boldsymbol{z}_t(p)\middle|\Delta_{\boldsymbol{\sigma}(r)}\right] = \sum_{t\geq t_0} \mathbb{E}\Big[\delta_t \boldsymbol{u}_t(p)\boldsymbol{z}_t(p)\Big|\Delta_{\boldsymbol{\sigma}(r)}\Big].$$

We now show that each term $\mathbb{E}[\delta_t \boldsymbol{u}_t(p)\boldsymbol{z}_t(p) \mid \Delta_{\boldsymbol{\sigma}(r)}]$ is zero for each t. Suppose that we further condition on all previous random decisions made by the algorithm before step size δ_t is chosen. In this case, the quantity $\boldsymbol{u}_t(p)\boldsymbol{z}_t(p)$ is completely determined and so δ_t is independent of $\boldsymbol{u}_t(p)\boldsymbol{z}_t(p)$. Moreover, the step size has mean zero, as shown in the proof of Lemma 2.7. Thus, for $t \geq t_0$,

$$\mathbb{E}[\delta_t \boldsymbol{u}_t(p)\boldsymbol{z}_t(p) \mid \delta_1 \dots \delta_{t-1}] = \boldsymbol{u}_t(p)\boldsymbol{z}_t(p) \cdot \mathbb{E}[\delta_t \mid \delta_1 \dots \delta_{t-1}] = 0$$

By the law of total expectation, it follows that the term $\mathbb{E}[\delta_t \boldsymbol{z}_t(p) \mid \Delta_{\boldsymbol{\sigma}(r)}]$ is zero for $t \geq t_0$. Thus, the second term in (A.2) is zero and so we have that

$$\mathbb{E}\left[\sum_{t\in L_r} \delta_t^2 \;\middle|\; \Delta_{\boldsymbol{\sigma}(r)}\right] = 1 - \boldsymbol{z}_{t_0}(p)^2 \leq 1,$$

where the inequality follows from $\boldsymbol{z}_{t_0}(p) \in (-1, 1)$. □

At this point, we are ready to prove the covariance bound.

Theorem 2.12*. *If all input vectors $\boldsymbol{b}_1 \dots \boldsymbol{b}_n$ have ℓ_2 norm at most one, then the covariance matrix of the vector of imbalances $\boldsymbol{Bz}$ is bounded in the Loewner order by the orthogonal projection onto the subspace spanned by the columns of $\boldsymbol{B}$:*

$$\operatorname{Cov}(\boldsymbol{Bz}) \preceq \boldsymbol{P} = \boldsymbol{B}\big(\boldsymbol{B}^\mathsf{T}\boldsymbol{B}\big)^\dagger \boldsymbol{B}^\mathsf{T},$$

where we recall that $\boldsymbol{A}^\dagger$ denotes the pseudoinverse of the matrix $\boldsymbol{A}$.

Proof. To prove the matrix inequality in the statement of the theorem, we seek to show that

$$\boldsymbol{v}^\mathsf{T} \operatorname{Cov}(\boldsymbol{Bz})\boldsymbol{v} \leq \boldsymbol{v}^\mathsf{T}\boldsymbol{P}\boldsymbol{v} \qquad \text{for all } \boldsymbol{v} \in \mathbb{R}^m$$

Using Lemma A.3 for the form of $\operatorname{Cov}(\boldsymbol{z})$ and linearity of expectation, we have that

$$\boldsymbol{v}^\mathsf{T} \operatorname{Cov}(\boldsymbol{Bz})\boldsymbol{v} = \boldsymbol{v}^\mathsf{T}\boldsymbol{B} \operatorname{Cov}(\boldsymbol{z})\boldsymbol{B}^\mathsf{T}\boldsymbol{v} = \boldsymbol{v}^\mathsf{T}\boldsymbol{B}\,\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \boldsymbol{u}_t\boldsymbol{u}_t^\mathsf{T}\right]\boldsymbol{B}^\mathsf{T}\boldsymbol{v} = \mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right].$$

Thus, we seek to show that for all $\boldsymbol{v} \in \mathbb{R}^m$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right] \leq \boldsymbol{v}^\mathsf{T} \boldsymbol{P} \boldsymbol{v}.$$

Next, we compute an upper bound on the quadratic forms in the sum. For each iteration t,

$$\begin{aligned}
\langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2 &= \left\langle \sum_{r=\ell_t}^{g_t} \big\langle \boldsymbol{w}_{\boldsymbol{\sigma}(i)}, \boldsymbol{b}_{p_t}\big\rangle \boldsymbol{w}_{\boldsymbol{\sigma}(i)}, \boldsymbol{v} \right\rangle^2 && \text{(Lemma A.2)} \\
&= \left(\sum_{r=\ell_t}^{g_t} \big\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t}\big\rangle \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle \right)^2 && \text{(linearity)} \\
&\leq \left(\sum_{r=\ell_t}^{g_t} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_{p_t}\rangle^2 \right) \left(\sum_{r=\ell_t}^{g_t} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \right) && \text{(Cauchy–Schwarz)} \\
&\leq \|\boldsymbol{b}_{p_t}\|^2 \cdot \left(\sum_{r=\ell_t}^{g_t} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \right) && (\boldsymbol{w}_{\boldsymbol{\sigma}(r)} \text{ are orthonormal}) \\
&\leq \left(\sum_{r=\ell_t}^{g_t} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \right). && (\text{by assumption, } \|\boldsymbol{b}_{p_t}\|^2 \leq 1)
\end{aligned}$$

Using this upper bound, we obtain an upper bound for the expected quantity of interest,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \left(\sum_{r=\ell_t}^{g_t} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \right)\right] && \text{(from above)} \\
&= \mathbb{E}\left[\sum_{r=1}^{n} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \sum_{t \in L_r} \delta_t^2\right] && \text{(rearranging terms)} \\
&= \sum_{r=1}^{n} \mathbb{E}\left[\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \sum_{t \in L_r} \delta_t^2\right] && \text{(linearity of expectation)}
\end{aligned}$$

We examine each of the terms in this sum. Fix a position r in the random reordering. Suppose that we further condition on $\Delta_{\boldsymbol{\sigma}(r)}$, which contains all random decisions made by the algorithm up to and including when unit $\boldsymbol{\sigma}(r)$ was decided by the algorithm. Under this conditioning, the vector $\boldsymbol{w}_{\boldsymbol{\sigma}(r)}$ is completely determined and so the quantity $\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2$ is also completely determined. In this way, the random term $\sum_{t \in L_r} \delta_t^2$ is conditionally independent of $\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2$ given $\Delta_{\boldsymbol{\sigma}(r)}$. Thus, we have

that

$$\mathbb{E}\left[\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \sum_{t\in L_r} \delta_t^2 \middle| \Delta_{\boldsymbol{\sigma}(r)}\right] = \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \cdot \mathbb{E}\left[\sum_{t\in L_r} \delta_t^2 \middle| \Delta_{\boldsymbol{\sigma}(r)}\right] \leq \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2,$$

where the equality is due to conditional independence and the inequality follows from Lemma A.4. Using iterated expectation, it follows that

$$\mathbb{E}\left[\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2 \sum_{t\in L_r} \delta_t^2\right] \leq \mathbb{E}\Big[\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2\Big].$$

Substituting this bound and using linearity of expectation yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \delta_t^2 \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle^2\right] \leq \sum_{r=1}^{n} \mathbb{E}\Big[\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2\Big] = \boldsymbol{v}^\intercal\, \mathbb{E}\left[\sum_{r=1}^{n} \boldsymbol{w}_{\boldsymbol{\sigma}(r)} \boldsymbol{w}_{\boldsymbol{\sigma}(r)}^\intercal\right] \boldsymbol{v} = \boldsymbol{v}^\intercal \boldsymbol{P} \boldsymbol{v} \ ,$$

where the last equality follows from the fact that the vectors $\boldsymbol{w}_{\boldsymbol{\sigma}(1)}, \boldsymbol{w}_{\boldsymbol{\sigma}(2)}, \ldots, \boldsymbol{w}_{\boldsymbol{\sigma}(n)}$ form an orthonormal basis for the span of input vectors, thus $\sum_{r=1}^{n} \boldsymbol{w}_{\boldsymbol{\sigma}(r)} \boldsymbol{w}_{\boldsymbol{\sigma}(r)}^\intercal = \boldsymbol{P}$ holds deterministically, regardless of the randomized re-ordering. □

A.1.4 Subgaussian bound (Theorem 2.19)

In this section, we prove an extended version of the subgaussian concentration inequality of Theorem 2.19. We begin by presenting the main technical inequality (Lemma A.5) which is stated in terms of operator monotonicity and proved using basic calculus. Next, we present Lemma A.6, which analyzes the behavior of the Gram–Schmidt Walk algorithm in one pivot phase using a backwards induction style argument. Finally, we prove the subgaussian concentration inequality by showing how we may repeatedly apply Lemma A.6.

The main technical inequality is stated in terms of operator monotonicity, which we briefly describe here. Let $\mathcal{D}$ be a set of n-by-n symmetric matrices. A real-valued matrix function $f : \mathcal{D} \to \mathbb{R}$ is said to be *operator monotone increasing* if

$$\boldsymbol{A}, \boldsymbol{B} \in \mathcal{D} \text{ with } \boldsymbol{A} \preceq \boldsymbol{B} \Rightarrow f(\boldsymbol{A}) \leq f(\boldsymbol{B}).$$

Intuitively, a real-valued matrix function f is monotone increasing if "larger" matrices (as determined by the Loewner order) are assigned larger values. We say that f is *operator monotone decreasing* if $\boldsymbol{A} \preceq \boldsymbol{B}$ implies instead that $f(\boldsymbol{A}) \geq f(\boldsymbol{B})$. Although there is a well developed theory of operator monotonicity, we use only very basic facts here which are mostly self contained. For more information on operator monotonicity, we refer readers to Chapter 5 of Bhatia (1997).

Lemma A.5. *For all $x \in [-1, 1]$ the function*

$$f_x \begin{pmatrix} \alpha & \eta \\ \eta & \beta \end{pmatrix} = \exp\left(-\frac{1}{2}\alpha\beta\right)\left[\frac{1+x}{2}\exp((1-x)\eta) + \frac{1-x}{2}\exp(-(1+x)\eta)\right]$$

is operator monotone decreasing over the set of 2-by-2 positive semidefinite matrices.

Proof. Operator monotonicity of a function $g : \mathcal{D} \to \mathbb{R}$ is preserved under composition with any monotone increasing $h : \mathbb{R} \to \mathbb{R}$. Using this and observing that f_x takes positive values for $x \in [-1, 1]$, we have that f_x is operator monotone decreasing if and only if $\log f_x$ is operator monotone decreasing. Moreover, a differentiable function $g : \mathcal{D} \to \mathbb{R}$ is operator monotone decreasing if and only if $-\nabla g(\boldsymbol{A})$ is positive semidefinite for all $\boldsymbol{A} \in \mathcal{D}$. The function f_x under consideration is differentiable and thus, to prove the lemma, it suffices to show that

$$-\nabla \log f_x \begin{pmatrix} \alpha & \eta \\ \eta & \beta \end{pmatrix}$$

is positive semidefinite when the 2-by-2 input matrix is positive semidefinite, i.e., $\alpha, \beta \geq 0$ and $\alpha\beta \geq \eta^2$.

We begin by defining the shorthand

$$\psi_x(\eta) = \log\left[\frac{1+x}{2}\exp((1-x)\eta) + \frac{1-x}{2}\exp(-(1+x)\eta)\right]$$

for the log of the bracketed term in the definition of f_x. Using this, we may write the function $\log f_x$ as

$$\log f_x \begin{pmatrix} \alpha & \eta \\ \eta & \beta \end{pmatrix} = \psi_x(\eta) - \frac{1}{2}\alpha\beta.$$

From the above expression, it is clear that $\partial_\alpha \log f_x = -\beta/2$, $\partial_\beta \log f_x = -\alpha/2$, and $\partial_\eta \log f_x = \partial_\eta \psi_x$. Thus, the matrix gradient may be computed:

$$-2\nabla \log f_x = \begin{pmatrix} \beta & -\partial_\eta \psi_x(\eta) \\ -\partial_\eta \psi_x(\eta) & \alpha \end{pmatrix}.$$

Recall that when computing the matrix gradient, we scale the off diagonals by $1/2$, as they appear twice in the trace inner product. We seek to show that the matrix above is positive semidefinite when the input matrix is positive semidefinite. Because the matrix above is 2-by-2, proving that it is positive semidefinite is equivalent to showing the three inequalities $\alpha, \beta \geq 0$ and $\alpha\beta \geq (\partial_\eta \psi_x(\eta))^2$. Because the input matrix is positive semidefinite, we already have that $\alpha, \beta \geq 0$. To show the final inequality, we show in the next part of the proof that $\eta^2 \geq (\partial_\eta \psi_x(\eta))^2$. Because the input matrix already satisfies $\alpha\beta \geq \eta^2$, this will imply the final inequality.

So for the final part of the proof, we focus on showing the inequality

$$(\partial_\eta \psi_x(\eta))^2 \leq \eta^2 \quad \text{for all } x \in [-1, 1].$$

To this end, we use an enveloping argument to show that $|\partial_\eta \psi_x(\eta)| \leq |\eta|$ for all $x \in [-1, 1]$. We begin by computing the first and second derivatives of $\psi_x(\eta)$. First, we rewrite the function $\psi_x(\eta)$ as

$$\begin{aligned}
\psi_x(\eta) &= \log\left[\frac{1+x}{2}\exp(1-x)\eta + \frac{1-x}{2}\exp -(1+x)\eta\right] \\
&= \log\left[\frac{1}{2}\big(e^{\eta - x\eta} + xe^{\eta - x\eta} + e^{-\eta - x\eta} - xe^{-\eta - x\eta}\big)\right] \\
&= \log\left[\frac{e^{-x\eta}}{2}(e^{\eta} + xe^{\eta} + e^{-\eta} - xe^{-\eta})\right] \\
&= \log\left[\frac{1}{2}(e^{\eta} + xe^{\eta} + e^{-\eta} - xe^{-\eta})\right] - x\eta \\
&= \log[\cosh(\eta) + x\sinh(\eta)] - x\eta.
\end{aligned}$$

Next, we compute the derivative $\partial_\eta \psi_x(\eta)$ by using chain rule and derivatives of log and hyperbolic trigonometric functions:

$$\partial_\eta \psi_x(\eta) = \frac{\sinh(\eta) + x\cosh(\eta)}{\cosh(\eta) + x\sinh(\eta)} - x.$$

Finally, we compute the second derivative of $\psi_x(\eta)$ using the above result, the quotient rule, and derivatives for the hyperbolic functions:

$$\partial_\eta^2 \psi_x(\eta) = 1 - \left(\frac{\sinh(\eta) + x\cosh(\eta)}{\cosh(\eta) + x\sinh(\eta)}\right)^2 = 1 - (\partial_\eta \psi_x(\eta) + x)^2.$$

We now establish the basis of our enveloping argument. That is, we show that the second derivative of $\psi_x(\eta)$ is bounded above and below by

$$0 \leq \partial_\eta^2 \psi_x(\eta) \leq 1 \quad \text{for all} \quad \eta \in \mathbb{R} \quad \text{and} \quad x \in [-1, 1].$$

The upper bound is immediate from the earlier expression, as $\partial_\eta^2 \psi_x(\eta) = 1 - (\partial_\eta \psi_x(\eta) +$

$x)^2 \leq 1$. The lower bound is a consequence of $x \in [-1, 1]$. To see this, observe that

$$\begin{aligned}
\partial_\eta^2 \psi_x(\eta) = 1 - \left(\frac{\sinh(\eta) + x\cosh(\eta)}{\cosh(\eta) + x\sinh(\eta)}\right)^2 &\geq 0 \\
&\Leftrightarrow (\cosh(\eta) + x\sinh(\eta))^2 \geq (\sinh(\eta) + x\cosh(\eta))^2 \\
&\Leftrightarrow \cosh^2(\eta) + x^2\sinh^2(\eta) \geq \sinh^2(\eta) + x^2\cosh^2(\eta) \\
&\Leftrightarrow \cosh^2(\eta) - \sinh^2(\eta) \geq x^2(\cosh^2(\eta) - \sinh^2(\eta)) \\
&\Leftrightarrow 1 \geq x^2
\end{aligned}$$

Now, we make our enveloping argument. First, we observe that $\partial_\eta \psi_x(0) = 0$. Next, for $\eta > 0$, we can bound the value of $\partial_\eta \psi_x(\eta)$ from above and below by

$$\begin{aligned}
\partial_\eta \psi_x(\eta) = \partial_\eta \psi_x(0) + \int_{y=0}^{\eta} \partial_\eta^2 \psi_x(y) dy &\leq 0 + \int_{y=0}^{\eta} 1 dy = \eta \\
\partial_\eta \psi_x(\eta) = \partial_\eta \psi_x(0) + \int_{y=0}^{\eta} \partial_\eta^2 \psi_x(y) dy &\geq 0 + \int_{y=0}^{\eta} 0 dy = 0.
\end{aligned}$$

Written together, these inequalities state that $0 \leq \partial_\eta \psi_x(\eta) \leq \eta$ for values $\eta \geq 0$. A similar enveloping argument shows that $-\eta \leq \partial_\eta \psi_x(\eta) \leq 0$ for values $\eta \leq 0$. Putting these two together, we have that $|\partial_\eta \psi_x(\eta)| \leq |\eta|$ for all $\eta \in \mathbb{R}$ and $x \in [-1, 1]$, as desired. □

Lemma A.6. *Let p be a unit that is chosen as the pivot and let Δ_p denote all random decisions made by the algorithm up until the beginning of pivot phase p. If $\|\boldsymbol{b}_p\| \leq 1$, then for all $\boldsymbol{v} \in \mathbb{R}^m$,*

$$\mathbb{E}\left[\exp\left(\sum_{t \in S_p} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle - \frac{1}{2}\|\boldsymbol{P}_p \boldsymbol{b}_p\|^2 \cdot \|\boldsymbol{P}_p \boldsymbol{v}\|^2\right) \,\middle|\, \Delta_p\right] \leq 1,$$

where S_p is the set of iterations for which p is the pivot.

Proof. Let t_p be the iteration at which p is first chosen to be the pivot. This iteration t_p is a deterministic quantity conditioned on Δ_p.

We begin by describing a convention which we adopt for the purposes of this analysis. Recall that the number of iterations in a pivot phase is generally a random quantity; however, the number of iterations in a pivot phase is at most n by Lemma 2.10. In fact, because $t_p - 1$ iterations have already occurred, the number of iterations in the pivot phase S_p is at most $n - t_p + 1$. For the purposes of this proof, we adopt a convention which deterministically fixes the number of iterations within the pivot phase to be $n - t_p + 1$. We adopt this convention because fixing the number of iterations in a pivot phase to be a deterministic quantity simplifies our backwards induction style argument. Once the pivot is frozen at iteration t, all remaining iterations of the pivot phase $s > t$ have step size zero, i.e. $\delta_s = 0$. In this

way, the fractional assignment is not updated in the remainder of the pivot phase after the pivot is frozen and thus this convention does not change the behavior of the algorithm. We emphasize again that this convention is for purposes of the current analysis and does not change the algorithm itself.

Using this convention and writing the iterations in the pivot phase as $S_p = \{t_p \dots n\}$, we seek to show that

$$\mathbb{E}\left[\exp\left(\sum_{t=t_p}^{n} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle - \frac{1}{2}\|\boldsymbol{P}_p \boldsymbol{b}_p\|^2 \cdot \|\boldsymbol{P}_p \boldsymbol{v}\|^2\right) \,\middle|\, \Delta_p\right] \leq 1. \tag{A.3}$$

All expectations in the remainder of the proof are conditioned on Δ_p and so we drop this notation.

We now rewrite the terms in the exponent by using the sequence of orthonormal basis vectors produced by the Gram–Schmidt orthogonalization process, as described in Section A.1.2. Suppose that the pivot unit has position $r = \boldsymbol{\sigma}^{-1}(p)$ in the reordering so that the $k+1$ vectors which are decided during this pivot phase appear in the ordering as

$$\boldsymbol{b}_{\boldsymbol{\sigma}(r-k)}, \boldsymbol{b}_{\boldsymbol{\sigma}(r-k+1)}, \dots \boldsymbol{b}_{\boldsymbol{\sigma}(r)},$$

where the pivot vector is the last in this re-ordering, i.e., $\boldsymbol{\sigma}(r) = p$, and so $\boldsymbol{b}_{\boldsymbol{\sigma}(r)} = \boldsymbol{b}_p$. The corresponding basis vectors produced by the Gram–Schmidt orthogonalization are

$$\boldsymbol{w}_{\boldsymbol{\sigma}(r-k)}, \boldsymbol{w}_{\boldsymbol{\sigma}(r-k+1)}, \dots \boldsymbol{w}_{\boldsymbol{\sigma}(r)}.$$

We now define a way to partition these reordering positions according to the iterations when they were decided. For each iteration $t = t_p, \dots n$ in this pivot phase, we define Q_t to be the reordering positions of the units that are frozen during the fractional assignment update in Line 12 during iteration t. By our convention, it may happen that $\delta_t = 0$ and in this case, $Q_t = \emptyset$. We also define $Q_{t_p-1} = \{g_p\} = \{\boldsymbol{\sigma}^{-1}(p)\}$, which is the re-ordering index of the pivot. We remark that this reordering position is deterministic given the conditioning Δ_p and the subscript $t_p - 1$ is chosen for notational convenience. Note that the reordering positions are determined in the order $Q_{t_p-1}, Q_{t_p}, \dots Q_n$ and this forms a partition of the reordering positions decided in this pivot phase.

Lemma A.2 shows that for each iteration t,

$$\boldsymbol{B}\boldsymbol{u}_t = \sum_{s=t_p-1}^{t-1} \sum_{r \in Q_s} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_p\rangle \boldsymbol{w}_{\boldsymbol{\sigma}(r)} \quad \text{and so} \quad \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle = \sum_{s=t_p-1}^{t-1} \sum_{r \in Q_s} \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_p\rangle \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle.$$

Recall that the projection matrix $\boldsymbol{P}_p$ is defined as

$$\boldsymbol{P}_p = \sum_{s=t_p-1}^{n} \sum_{r \in Q_s} \boldsymbol{w}_{\boldsymbol{\sigma}(r)} \boldsymbol{w}_{\boldsymbol{\sigma}(r)}^{\mathsf{T}}$$

and thus we have that

$$\|\boldsymbol{P}_p\boldsymbol{b}_p\|^2 = \sum_{s=t_p-1}^{n}\sum_{r\in Q_s}\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_p\rangle^2 \quad \text{and} \quad \|\boldsymbol{P}_p\boldsymbol{v}\|^2 = \sum_{s=t_p-1}^{n}\sum_{r\in Q_s}\langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle^2$$

For notational convenience, for each reordering position r, let $\alpha_r = \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{b}_p\rangle$ and $\beta_r = \langle \boldsymbol{w}_{\boldsymbol{\sigma}(r)}, \boldsymbol{v}\rangle$.

Substituting these terms into (A.3), we have that the desired inequality may be written as

$$\mathbb{E}\left[\exp\left(\sum_{t=t_p}^{n}\delta_t\sum_{s=t_p-1}^{t-1}\sum_{r\in Q_s}\alpha_r\beta_r - \frac{1}{2}\Big(\sum_{s=t_p-1}^{n}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{n}\sum_{r\in Q_s}\beta_r^2\Big)\right)\right] \le 1.$$

We will prove this inequality using a backwards induction style argument. We use the main technical inequality of Lemma A.5 to show that, conditioned on the first $n-1$ iterations, the expectation above is maximized when $\alpha_r = \beta_r = 0$ for all $r \in Q_n$. In some sense, this is identifying the worst-case values that $\{(\alpha_r, \beta_r) : r \in Q_n\}$ may take. We then continue backwards and show that given the values of $\{(\alpha_r, \beta_r) : r \in Q_t\}$ for $t < R$, the values of $\{(\alpha_r, \beta_r) : r \in \cup_{s=R}^{n} Q_s\}$ which maximize the expectation are $\alpha_r = \beta_r = 0$.

We now proceed more formally. For each $R = 0, 1, \ldots n$, we define the quantity

$$g(R) = \mathbb{E}\left[\exp\left(\Big(\sum_{t=t_p}^{n}\delta_t\sum_{s=t_p-1}^{\min\{R,t-1\}}\sum_{r\in Q_s}\alpha_r\beta_r\Big) - \frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)\right)\right]$$

Note that $g(R)$ is similar to the expectation we are interested in bounding, except that $\alpha_r = \beta_r = 0$ for all $r \in \cup_{s>R} Q_s$. Note that $g(n)$ is exactly the expectation that we seek to upper bound by 1. We prove this upper bound by establishing the following chain of inequalities

$$g(n) \le g(n-1) \le \cdots \le g(t_p) \le 1.$$

We prove this chain of inequalities in three steps. The first step is to establish that $g(n) \le g(n-1)$. This inequality is the simplest one to establish because it follows directly from the definition of $g(R)$. In particular, observe that the term $\sum_{t=t_p}^{n}\delta_t\sum_{s=t_p-1}^{\min\{R,t-1\}}\sum_{r\in Q_s}\alpha_r\beta_r$ is the same for $R = n$ and $R = n-1$, while the term $\frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)$ is larger for $R = n$ than for $R = n-1$. Thus, $g(n) \le g(n-1)$.

We now show the second chunk of inequalities: $g(R) \le g(R-1)$ for $t_p < R \le n-1$. Before continuing, we show how to use the main technical inequality (Lemma A.5)

to prove that for all R in this range,

$$\mathbb{E}\left[\exp\left(\Big(\sum_{t=R+1}^{n}\delta_t\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r\beta_r\Big)-\frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)\right)\,\middle|\,\Delta_R\right] \tag{A.4}$$

$$\leq\mathbb{E}\left[\exp\left(\Big(\sum_{t=R+1}^{n}\delta_t\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\alpha_r\beta_r\Big)-\frac{1}{2}\Big(\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\beta_r^2\Big)\right)\,\middle|\,\Delta_R\right],$$

where Δ_R denotes the step sizes, $\delta_{t_p},\delta_{t_p+1},\dots\delta_R$, in addition to the previous randomness in the algorithm denoted by Δ_p. Under this conditioning, the values of $\{(\alpha_r,\beta_r):r\in\cup_{s=t_p-1}^{R}Q_s\}$ are decided and the only random quantity in the expression above is $\sum_{t=R+1}^{n}\delta_t$. We claim that this random variable is precisely

$$\sum_{t=R+1}^{n}\delta_t=\begin{cases}1-\boldsymbol{z}_{R+1}(p) & \text{with probability } (1+\boldsymbol{z}_{R+1}(p))/2\\ -(1+\boldsymbol{z}_{R+1}(p)) & \text{with probability } (1-\boldsymbol{z}_{R+1}(p))/2\end{cases}$$

To see this, observe that because the step direction satisfies $u_t(p)=1$ in the pivot phase p and the update procedure is $\boldsymbol{z}_{t+1}\leftarrow\boldsymbol{z}_t+\delta_t\boldsymbol{u}_t$,

$$\boldsymbol{z}_n(p)=\sum_{t=R+1}^{n}\delta_t u_t(p)+\boldsymbol{z}_{R+1}(p)=\sum_{t=R+1}^{n}\delta_t+\boldsymbol{z}_{R+1}(p)\quad\text{and thus}\quad\sum_{t=R+1}^{n}\delta_t=\boldsymbol{z}_n(p)-\boldsymbol{z}_{R+1}(p).$$

Because $\boldsymbol{z}_n(p)$ takes values ±1, we have that the sum $\sum_{t=R+1}^{n}\delta_t$ only takes two values. Moreover, because all step sizes have mean zero, we have that $\mathbb{E}[\sum_{t=R+1}^{n}\delta_t]=0$. This determines the probabilities of each of the two values.

Because we know exactly the distribution of the random sum $\sum_{t=R+1}^{n}\delta_t$, we may derive the expectation in the left hand side of (A.4) exactly as

$$\frac{1+\boldsymbol{z}_{R+1}(p)}{2}\exp\left((1-\boldsymbol{z}_{R+1}(p))\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r\beta_r-\frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)\right) \tag{A.5}$$

$$+\frac{1-\boldsymbol{z}_{R+1}(p)}{2}\exp\left(-(1+\boldsymbol{z}_{R+1}(p))\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r\beta_r-\frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)\right)$$

We now demonstrate how this expectation may be recognized as the matrix function appearing in Lemma A.5. Let $\boldsymbol{A}$ and $\boldsymbol{A}_R$ be the 2-by-2 matrices given by

$$\boldsymbol{A}=\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\begin{pmatrix}\alpha_r^2 & \alpha_r\beta_r\\ \alpha_r\beta_r & \beta_r^2\end{pmatrix},\quad \boldsymbol{A}_R=\sum_{r\in Q_R}\begin{pmatrix}\alpha_r^2 & \alpha_r\beta_r\\ \alpha_r\beta_r & \beta_r^2\end{pmatrix}.$$

These matrices are the sum of 2-by-2 positive semidefinite matrices and so they are themselves positive semidefinite. Recall that the matrix function in Lemma A.5 is defined for $x \in [-1, 1]$ as

$$\begin{aligned} f_x\begin{pmatrix}\alpha & \eta \\ \eta & \beta\end{pmatrix} &= e^{-\frac{1}{2}\alpha\beta}\left[\frac{1+x}{2}\exp((1-x)\eta) + \frac{1-x}{2}\exp(-(1+x)\eta)\right] \\ &= \frac{1+x}{2}\exp\Big((1-x)\eta - \frac{1}{2}\alpha\beta\Big) + \frac{1-x}{2}\exp\Big(-(1+x)\eta - \frac{1}{2}\alpha\beta\Big). \end{aligned}$$

Observe that the expectation in (A.5) is equal to $f_{\boldsymbol{z}_R(p)}(\boldsymbol{A} + \boldsymbol{A}_R)$. By Lemma A.5, the function is operator monotone decreasing over positive semidefinite matrices so that

$$f_{\boldsymbol{z}_R(p)}(\boldsymbol{A} + \boldsymbol{A}_R) \leq f_{\boldsymbol{z}_R(p)}(\boldsymbol{A}).$$

The proof of inequality (A.4) is completed by observing that $f_{\boldsymbol{z}_R(p)}(\boldsymbol{A})$ is equal to the expectation on the right hand side of (A.4).

Now we are ready to show that $g(R) \leq g(R-1)$ for $t_p < R \leq n-1$. For notational convenience, we define

$$X_R = \exp\left(\sum_{t=t_p}^{R}\delta_t \sum_{s=t_p-1}^{t-1}\alpha_r\beta_r\right).$$

By rearranging terms, applying iterated expectations, and using the inequality (A.4), we have that

$$\begin{aligned} &g(R) \\ &= \mathbb{E}\left[\exp\left(\sum_{t=t_p}^{n}\delta_t \sum_{s=t_p-1}^{\min\{R,t-1\}}\sum_{r\in Q_s}\alpha_r\beta_r - \frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)\right)\right] \\ &= \mathbb{E}\left[X_R\cdot\exp\left(\sum_{t=R+1}^{n}\delta_t \sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r\beta_r - \frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)\right)\right] \\ &= \mathbb{E}\left[X_R\cdot\mathbb{E}\left[\exp\left(\sum_{t=R+1}^{n}\delta_t \sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r\beta_r - \frac{1}{2}\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R}\sum_{r\in Q_s}\beta_r^2\Big)\right)\,\middle|\,\Delta_R\right]\right] \\ &\leq \mathbb{E}\left[X_R\cdot\mathbb{E}\left[\exp\left(\sum_{t=R+1}^{n}\delta_t \sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\alpha_r\beta_r - \frac{1}{2}\Big(\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\beta_r^2\Big)\right)\,\middle|\,\Delta_R\right]\right] \\ &= \mathbb{E}\left[\exp\left(\sum_{t=t_p}^{n}\delta_t \sum_{s=t_p-1}^{\min\{R-1,t-1\}}\sum_{r\in Q_s}\alpha_r\beta_r - \frac{1}{2}\Big(\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\alpha_r^2\Big)\cdot\Big(\sum_{s=t_p-1}^{R-1}\sum_{r\in Q_s}\beta_r^2\Big)\right)\right] \\ &= g(R-1) \end{aligned}$$

This establishes the chain of inequalities

$$g(n) \leq g(n-1) \leq \cdots \leq g(t_p).$$

Establishing that $g(t_p) \leq 1$ may be done via a similar application of the operator monotonicity result of Lemma A.5. In particular,

$$\begin{aligned} g(t_p) &= \mathbb{E}\left[\exp\left(\Big(\sum_{t=t_p}^{n} \delta_t\Big)\langle \boldsymbol{w}_p, \boldsymbol{b}_p\rangle\langle \boldsymbol{w}_p, \boldsymbol{v}\rangle - \frac{1}{2}\langle \boldsymbol{w}_p, \boldsymbol{b}_p\rangle^2\langle \boldsymbol{w}_p, \boldsymbol{v}\rangle^2\right)\right] \\ &= f_{\boldsymbol{z}_{t_p}(p)}\left(\begin{bmatrix} \langle \boldsymbol{w}_p, \boldsymbol{b}_p\rangle^2 & \langle \boldsymbol{w}_p, \boldsymbol{b}_p\rangle \\ \langle \boldsymbol{w}_p, \boldsymbol{b}_p\rangle & \langle \boldsymbol{w}_p, \boldsymbol{v}\rangle^2 \end{bmatrix}\right) \\ &\leq f_{\boldsymbol{z}_{t_p}(p)}(\mathbf{0}) = 1. \end{aligned}$$

□

We now present the proof of the subgaussian concentration result.

Theorem 2.19*. *If the input vectors $\boldsymbol{b}_1 \ldots \boldsymbol{b}_n$ all have ℓ_2 norm at most 1, then the Gram–Schmidt Walk algorithm returns an assignment vector $\boldsymbol{z}$ so that the vector of imbalances $\boldsymbol{B}\boldsymbol{z}$ is subgaussian with variance parameter $\sigma^2 = 1$:*

$$\mathbb{E}\Big[\exp\Big(\langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{v}\rangle - \langle \mathbb{E}[\boldsymbol{B}\boldsymbol{z}], \boldsymbol{v}\rangle\Big)\Big] \leq \exp\big(\|\boldsymbol{v}\|^2/2\big) \qquad \textit{for all} \qquad \boldsymbol{v} \in \mathbb{R}^{n+d}.$$

Proof. We prove the stronger inequality

$$\mathbb{E}\Big[\exp\Big(\langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{v}\rangle - \langle \mathbb{E}[\boldsymbol{B}\boldsymbol{z}], \boldsymbol{v}\rangle\Big)\Big] \leq \mathbb{E}\Big[\exp\Big(\frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{P}_i\boldsymbol{b}_i\|^2\|\boldsymbol{P}_i\boldsymbol{v}\|^2\Big)\Big] \quad \text{for all} \quad \boldsymbol{v} \in \mathbb{R}^m. \tag{A.6}$$

To see that inequality (A.6) is stronger, we use the contractive property of projection matrices and the assumption that all input vectors have ℓ_2 norm at most 1 to show

$$\sum_{i=1}^{n}\|\boldsymbol{P}_i\boldsymbol{b}_i\|^2\|\boldsymbol{P}_i\boldsymbol{v}\|^2 \leq \sum_{i=1}^{n}\|\boldsymbol{b}_i\|^2\|\boldsymbol{P}_i\boldsymbol{v}\|^2 \leq \sum_{i=1}^{n}\|\boldsymbol{P}_i\boldsymbol{v}\|^2 = \|\boldsymbol{P}\boldsymbol{v}\|^2 \leq \|\boldsymbol{v}\|^2.$$

This shows that inequality (A.6) implies the inequality in the statement of the theorem.

We now rearrange and substitute terms in (A.6) to obtain a form that we will work with during the remainder of the proof. By dividing both sides of (A.6) by the right hand side, we obtain an equivalent expression of the inequality:

$$\mathbb{E}\Big[\exp\Big(\langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{v}\rangle - \langle \mathbb{E}[\boldsymbol{B}\boldsymbol{z}], \boldsymbol{v}\rangle - \frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{P}_i\boldsymbol{b}_i\|^2\|\boldsymbol{P}_i\boldsymbol{v}\|^2\Big)\Big] \leq 1 \quad \text{for all} \quad \boldsymbol{v} \in \mathbb{R}^m.$$

At this point, we drop the "for all $\boldsymbol{v} \in \mathbb{R}^m$" qualifier and assume that an arbitrary $\boldsymbol{v} \in \mathbb{R}^m$ is given. We re-write the quantity $\langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{v}\rangle - \langle \mathbb{E}[\boldsymbol{B}\boldsymbol{z}], \boldsymbol{v}\rangle$ in terms of the

fractional updates in the algorithm:

$$\langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{v}\rangle = \Big\langle \boldsymbol{B}\Big(\sum_{t=1}^{T} \delta_t \boldsymbol{u}_t + \boldsymbol{z}_1\Big), \boldsymbol{v}\Big\rangle = \sum_{t=1}^{T} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle + \langle \boldsymbol{B}\boldsymbol{z}_1, \boldsymbol{v}\rangle = \sum_{i=1}^{n}\sum_{t\in S_i} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle + \langle \boldsymbol{B}\boldsymbol{z}_1, \boldsymbol{v}\rangle.$$

Note that by the martingale property of the fractional updates (Lemma 2.7), $\mathbb{E}[\boldsymbol{z}] = \boldsymbol{z}_1$. Thus,

$$\langle \mathbb{E}[\boldsymbol{B}\boldsymbol{z}], \boldsymbol{v}\rangle = \langle \boldsymbol{B}\,\mathbb{E}[\boldsymbol{z}], \boldsymbol{v}\rangle = \langle \boldsymbol{B}\boldsymbol{z}_1, \boldsymbol{v}\rangle$$

and so the difference is given by

$$\langle \boldsymbol{B}\boldsymbol{z}, \boldsymbol{v}\rangle - \langle \mathbb{E}[\boldsymbol{B}\boldsymbol{z}], \boldsymbol{v}\rangle = \sum_{i=1}^{n}\sum_{t\in S_i} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle.$$

Using this expression for the difference, we may write the desired inequality, which features a sum over units in the exponent, as follows:

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^{n}\Big(\sum_{t\in S_i} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle - \frac{1}{2}\|\boldsymbol{P}_i \boldsymbol{b}_i\|^2 \|\boldsymbol{P}_i \boldsymbol{v}\|^2\Big)\right)\right] \le 1.$$

A unit $i \in [n]$ which is not chosen as the pivot does not contribute to this sum because the corresponding pivot phase S_i is empty and the projection matrix $\boldsymbol{P}_i$ is the zero. Thus, we may write the sum over units which are chosen as the pivot. We denote the sequence of pivot units as $p_1, p_2, \dots p_k$ where the subscripts denote the order in which the pivots are chosen by the algorithm. We seek to show that

$$\mathbb{E}\left[\exp\left(\sum_{j=1}^{k}\Big(\sum_{t\in S_{p_j}} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle - \frac{1}{2}\|\boldsymbol{P}_{p_j} \boldsymbol{b}_{p_j}\|^2 \|\boldsymbol{P}_{p_j} \boldsymbol{v}\|^2\Big)\right)\right] \le 1.$$

To this end, we define the sequence of random variables $X_1, X_2, \dots X_k$ by

$$X_j = \sum_{t\in S_{p_j}} \delta_t \langle \boldsymbol{B}\boldsymbol{u}_t, \boldsymbol{v}\rangle - \frac{1}{2}\|\boldsymbol{P}_{p_j} \boldsymbol{b}_{p_j}\|^2 \|\boldsymbol{P}_{p_j} \boldsymbol{v}\|^2,$$

where each X_j corresponds to the jth pivot that was chosen by the algorithm.[1] We show that $\mathbb{E}[\exp(\sum_{j=1}^{k} X_j)] \le 1$ by proving the chain of inequalities

$$\mathbb{E}\Big[\exp\Big(\sum_{j=1}^{k} X_j\Big)\Big] \le \mathbb{E}\Big[\exp\Big(\sum_{j=1}^{k-1} X_j\Big)\Big] \le \cdots \le \mathbb{E}[\exp(X_1)] \le \mathbb{E}[\exp(0)] = 1.$$

[1] In the proof sketch in Chapter 2, we used terms D_i which did not incorporate the projection $\|\boldsymbol{P}_{p_j} \boldsymbol{b}_{p_j}\|^2$, so $X_i \ge D_i$. By incorporating the projection terms in this full proof, we more clearly see the stronger inequality (A.6) that is being proven. This highlights that the subgaussian bound will be loose when $\|\boldsymbol{P}_{p_j} \boldsymbol{b}_{p_j}\|^2 \le 1$ is a loose inequality.

Consider some $1 \leq \ell \leq k$. Let Δ_ℓ be all random decisions made by the algorithm up until the beginning of pivot phase ℓ. Then observe that

$$\begin{aligned}
\mathbb{E}\Big[\exp\Big(\sum_{j=1}^{\ell} X_j\Big)\Big] &= \mathbb{E}\Big[\exp\Big(\sum_{j=1}^{\ell-1} X_j\Big) \cdot \exp(X_\ell)\Big] && \text{(property of exponential)} \\
&= \mathbb{E}\Big[\exp\Big(\sum_{j=1}^{\ell-1} X_j\Big) \cdot \mathbb{E}[\exp(X_\ell) \mid \Delta_\ell]\Big] && \text{(iterated expectations)} \\
&\leq \mathbb{E}\Big[\exp\Big(\sum_{j=1}^{\ell-1} X_j\Big)\Big], && \text{(by Lemma A.6)}
\end{aligned}$$

which completes the induction. □

A.1.5 Extending the analysis to the GSW-Design

In this section, we demonstrate that our analysis of the Gram–Schmidt Walk algorithm extends to the GSW-Design. The main difference between the Gram–Schmidt Walk algorithm and the GSW-Design are the construction of input vectors and the randomized pivoting rule. The randomized pivoting rule in the design is inconsequential to the theorems proved in this section. The purpose of the randomized pivoting rule is to allow us to prove that the second-order assignment probabilities are bounded away from zero, which we need in order to estimate the ridge loss, as discussed in Section 2.6.2.

We remark that the GSW-Design presented in Section 4.6 may be implemented as follows:

1. Construct the $(n+d)$-dimensional augmented covariate vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \dots \boldsymbol{b}_n$ as

$$\boldsymbol{b}_i = \begin{bmatrix} \sqrt{\phi}\boldsymbol{e}_i \\ \xi^{-1}\sqrt{1-\phi}\boldsymbol{x}_i \end{bmatrix},$$

   where $\boldsymbol{e}_i$ is the n-dimensional ith standard basis vector and $\xi = \max_{i\in[n]}\|\boldsymbol{x}_i\|$.

2. Permute the order of the input vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \dots \boldsymbol{b}_n$ with a uniformly random permutation.

3. Run the Gram–Schmidt Walk (Algorithm 5) with permuted input vectors and initial fractional assignment $\boldsymbol{z}_1 = \boldsymbol{0}$ to produce assignment vector $\boldsymbol{z}$.

The key idea behind the equivalence of these descriptions is that the method of uniformly permuting input vectors then deterministically choosing largest indexed alive unit as pivot (as presented here) produces the same distribution as choosing pivots uniformly from the set of alive units (as presented in Section 4.6). To see this equivalence, begin by considering the first iteration: the largest index in a uniformly

permuted list of units is uniform over all units. This means that the first pivot chosen by the two methods has the same distribution. Moreover, the construction of step direction and step size does not depend on the index of the units. In this way, a similar argument shows that these methods of selecting the pivot are equivalent: the largest index in a uniformly permuted list of alive units is uniform over all alive units. Thus, the two random pivot sampling schemes are equivalent.

Due to this equivalence, we may analyze the GSW-Design by applying the analysis in this section. Because the covariance bound (Theorem 2.12*) and the subgaussian concentration (Theorem 2.19*) hold for all orderings of the input vectors, they hold for any distribution over the orderings of the input vectors. In particular, they hold for the uniform distribution over orderings of the input vectors and so they apply to the GSW-Design.

Finally, we remark that the augmented covariate vectors constructed in the GSW-Design satisfy the condition that each of their ℓ_2 norms is at most one. This norm condition is a scaling requirement in order to make the covariance and subgaussian bounds in Theorem 2.12 and Theorem 2.19, respectively. To see that the norm condition holds, observe that

$$\|\boldsymbol{b}_i\|^2 = \left\|\sqrt{\phi}\boldsymbol{e}_i\right\|^2 + \left\|\xi^{-1}\sqrt{1-\phi}\boldsymbol{x}_i\right\|^2 = \phi + (1-\phi)\big(\xi^{-1}\|\boldsymbol{x}_i\|\big)^2 \leq \phi + (1-\phi) = 1,$$

where the inequality follows from the definition $\xi = \max_{i\in[n]}\|\boldsymbol{x}_i\|$.

Taken together, this shows that Theorems 2.12 and 2.19 in Chapter 2 follow from Theorems 2.12* and 2.19* in this supplement.

A.1.6 Non-uniform treatment probabilities

The GSW-Design can be extended to allow arbitrary assignment probabilities. We achieve this by changing the initial fractional assignments of the algorithm. The experimenter provides a parameter vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n) \in (0,1)^n$ specifying the desired first-order assignment probability for each unit. The first step of the algorithm in Section 4.6 is then modified so that $\boldsymbol{z}_1 \leftarrow 2\boldsymbol{\pi} - \mathbf{1}$. The following corollary is a direct consequence of the martingale property of the fractional updates, in the same fashion as Corollary 2.8.

Corollary A.7. *Under the non-uniform Gram–Schmidt Walk design,*

$$\Pr(z_i = 1) = \pi_i \qquad \textit{for all} \qquad i \in [n].$$

The properties of the original version of the design can be extended to the non-uniform version. To do so, we redefine the vector $\boldsymbol{\mu}$ as

$$\widetilde{\boldsymbol{\mu}} = \left(\frac{a_1}{4\pi_1} + \frac{b_1}{4(1-\pi_1)}, \ldots, \frac{a_n}{4\pi_n} + \frac{b_n}{4(1-\pi_n)}\right).$$

In this vector, each potential outcome is weighted by the probability that it is observed. If $\boldsymbol{\pi} = 0.5 \times \mathbf{1}$, then $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$, which replicates the uniform version of the design. The mean squared error of the Horvitz–Thompson estimator can now be expressed as

$$\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] = \frac{1}{n^2} \widetilde{\boldsymbol{\mu}}^\intercal \operatorname{Cov}(\boldsymbol{z}) \widetilde{\boldsymbol{\mu}}.$$

This extends Lemma 2.2 to any experimental design with non-deterministic assignments. In particular, Theorems 2.12 and 2.19 hold for the non-uniform version of the design, so all properties that follow from these theorems also apply to the extended version when $\widetilde{\boldsymbol{\mu}}$ is substituted for $\boldsymbol{\mu}$.

A.2 Fast Implementation of the GSW-Design

The most computationally intensive aspect of the Gram–Schmidt Walk is the computation of the step direction $\boldsymbol{u}_t$. Although it is defined as the solution to an optimization problem, it may be obtained efficiently by solving a system of linear equations. Computational speed ups may be obtained by pre-computing and maintaining a certain matrix factorization, decreasing the cost of repeated linear system solves at each iteration. In this section, we provide details of such an efficient implementation.

A.2.1 Derivation of the step direction

Recall that at each iteration t, the step direction $\boldsymbol{u}_t$ is defined as the vector which has coordinates $\boldsymbol{u}_t(i) = 0$ for $i \notin \mathcal{A}_t$, coordinate $\boldsymbol{u}_t(p_t) = 1$ for the pivot unit p_t, and the remaining coordinates are the solution to

$$\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t) = \arg\min_{\boldsymbol{u}} \| \boldsymbol{b}_{p_t} + \sum_{i \notin \mathcal{A}_t \setminus p_t} \boldsymbol{u}(i) \boldsymbol{b}_i \|^2 \ .$$

The minimization above is a least squares problem and the solution may be obtained by solving a system of linear equations. Let k be the number of units which are alive and not the pivot, i.e., $k = |\mathcal{A}_t \setminus p_t|$, and let $\boldsymbol{B}_t$ be the $(n+d)$-by-k matrix with columns $\boldsymbol{b}_i$ for $i \in \mathcal{A}_t \setminus p_t$. As the augmented covariate vectors are linearly independent, the coordinates $\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t)$ that minimize the quantity $\| \boldsymbol{b}_{p_t} + \boldsymbol{B}_t \boldsymbol{u}_t(\mathcal{A}_t \setminus p_t) \|^2$ are given by the normal equations

$$\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t) = -\big(\boldsymbol{B}_t^\intercal \boldsymbol{B}_t\big)^{-1} \boldsymbol{B}_t^\intercal \boldsymbol{b}_{p_t} \ .$$

Let $\boldsymbol{X}_t$ denote the row-submatrix of $\boldsymbol{X}$ with rows $\mathcal{A}_t \setminus p_t$. Using our specific form of $\boldsymbol{B}$, and by direct calculation and application of the Woodbury identity lemma, we obtain that

$$\big(\boldsymbol{B}_t^\intercal \boldsymbol{B}_t\big)^{-1} = \big(\phi \boldsymbol{I}_k + \xi^{-2}(1-\phi) \boldsymbol{X}_t \boldsymbol{X}_t^\intercal\big)^{-1} = \phi^{-1} \left[\boldsymbol{I}_k - \boldsymbol{X}_t \left(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \frac{\xi^2 \phi}{1-\phi} \boldsymbol{I}_d \right)^{-1} \boldsymbol{X}_t^\intercal \right] \ .$$

By again using our specific form of input matrix $\boldsymbol{B}$, a direct calculation yields that

$$\boldsymbol{B}_t^\intercal \boldsymbol{b}_{p_t} = \xi^{-2}(1-\phi)\boldsymbol{X}_t \boldsymbol{x}_{p_t} \ .$$

Thus, we obtain a form for the relevant coordinates in the update direction vector $\boldsymbol{u}_t$

$$\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t) = -\left(\frac{1-\phi}{\xi^2\phi}\right) \underbrace{\boldsymbol{X}_t}_{n\times d} \left[\boldsymbol{x}_{p_t} - \underbrace{\left(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \frac{\xi^2\phi}{1-\phi}\boldsymbol{I}_d\right)^{-1}}_{d\times d} \underbrace{\boldsymbol{X}_t^\intercal \boldsymbol{X}_t}_{d\times d} \boldsymbol{x}_{p_t} \right] \ , \quad \text{(A.7)}$$

which involves smaller matrices of size $d \times d$, rather than $n \times n$. In the next few paragraphs, we show how computing and maintaining factorizations of these smaller matrices results in faster computations of the step direction $\boldsymbol{u}_t$. We are chiefly concerned with computing and maintaining a factorization of the matrix $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d)$. We describe an implementation which uses the Cholesky factorization, although there are several appropriate alternatives.

A.2.2 Cholesky factorizations

Here, we briefly review Cholesky factorizations and their computational properties. The *Cholesky factorizatio*n of an n-by-n symmetric positive definite matrix $\boldsymbol{A}$ is the unique factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\intercal$, where $\boldsymbol{L}$ is lower triangular. Given the matrix $\boldsymbol{A}$, the matrix $\boldsymbol{L}$ may be obtained using $\mathcal{O}(n^3)$ arithmetic operations. Once the Cholesky factorization $\boldsymbol{L}$ is obtained, solutions $\boldsymbol{x}$ to the system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ may be computed using $\mathcal{O}(n^2)$ arithmetic operations by using a forward-backward algorithm which leverages the triangular structure of $\boldsymbol{L}$. In general, solving systems of linear equations takes $\mathcal{O}(n^3)$ arithmetic operations[2] and so if many linear system solves are required, then computing the factorization and using the faster forward-backward algorithm yields computational speed-ups. Suppose that $\boldsymbol{A}$ is a positive definite matrix with Cholesky factorization $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\intercal$ and that the rank-1 updated matrix $\boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^\intercal$ has Cholesky factorization $\boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^\intercal = \boldsymbol{L}_+\boldsymbol{L}_+^\intercal$. Given the original factorization $\boldsymbol{L}$ and the vector $\boldsymbol{v}$, the updated factorization $\boldsymbol{L}_+$ may be computed using $\mathcal{O}(n^2)$ arithmetic computations, without extra memory allocation. Updating in this way is a much more efficient way to maintain the factorization than explicitly computing $\boldsymbol{A} + \boldsymbol{v}\boldsymbol{v}^\intercal$ and its factorization directly. The same technique may be used for rank-1 downdates $\boldsymbol{A} - \boldsymbol{v}\boldsymbol{v}^\intercal$ when the updated matrix remains positive definite. For more details, see Stewart (1998); Trefethen and Bau (1997).

[2]While there are algorithms based on fast matrix multiplication that are asymptotically faster, they do not meaningfully change this discussion for realistic values of n.

A.2.3 Computing and maintaining factorizations

Before the first pivot is chosen, we have that $\boldsymbol{X}_t = \boldsymbol{X}$, as no rows of $\boldsymbol{X}$ have been decided. Thus, we compute $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d)$ directly and then compute a Cholesky factorization. Computing the matrix directly requires $\mathcal{O}(nd^2)$ time and computing the factorization requires $\mathcal{O}(d^3)$ time. Each time a variable $i \in [n]$ is frozen or chosen as the pivot, the set $\mathcal{A}_t \setminus p_t$ is updated and so we must update the factorization $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d)$. The update consists of removing the row vector $\boldsymbol{x}_i$ from $\boldsymbol{X}_t$. One can see that this corresponds to a rank-1 downdate to the entire matrix $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d)$. Rank-1 downdates to a Cholesky factorization may be computed in-place, using $\mathcal{O}(d^2)$ arithmetic operations. Because there will be at most n rank-1 updates to this factorization, the total update cost is $\mathcal{O}(nd^2)$ arithmetic operations. Thus, the total computational cost of maintaining this Cholesky factorization is $\mathcal{O}(nd^2)$ arithmetic operations and $\mathcal{O}(d^2)$ memory.

A.2.4 Computing step directions

Assume that at each iteration, we have a Choleksy factorization of the matrix $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d)$. By (A.7), we can solve for the relevant coordinates in the step direction $\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t)$ using the following three computations:

1. $\boldsymbol{a}_t^{(1)} = \boldsymbol{X}_t^\intercal \boldsymbol{X}_t \boldsymbol{x}_{p_t}$
2. $\boldsymbol{a}_t^{(2)} = \left(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d\right)^{-1}\boldsymbol{a}_t^{(1)}$
3. $\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t) = -\xi^{-2}\phi^{-1}(1-\phi)\boldsymbol{X}_t\left(\boldsymbol{x}_{p_t} - \boldsymbol{a}_t^{(2)}\right)$

If the matrix $\boldsymbol{X}_t^\intercal \boldsymbol{X}_t$ is explicitly available at the beginning of each iteration, then computing $\boldsymbol{a}_t^{(1)}$ can be done in $\mathcal{O}(d^2)$ time by matrix-vector multiplication. While it is possible to maintain $\boldsymbol{X}_t^\intercal \boldsymbol{X}_t$ explicitly, it requires an extra $\mathcal{O}(d^2)$ memory. On the other hand, if $\boldsymbol{X}_t^\intercal \boldsymbol{X}_t$ is not explicitly available, then $\boldsymbol{a}_t^{(1)}$ may be obtained from a factorization of $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d)$, as

$$\boldsymbol{a}_t^{(1)} = \left(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \frac{\xi^2\phi}{1-\phi}\boldsymbol{I}_d\right)\boldsymbol{x}_{p_t} - \left(\frac{\xi^2\phi}{1-\phi}\right)\boldsymbol{x}_{p_t} \ ,$$

which saves $\mathcal{O}(d^2)$ memory and incurs only a slightly larger arithmetic cost of $\mathcal{O}(d^2 + d)$. Next, one may compute $\boldsymbol{a}_t^{(2)}$ using $\mathcal{O}(d^2)$ arithmetic operations via a forward-backward solver on the Cholesky factorization. Finally, computing $\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t)$ may be done in $\mathcal{O}(nd)$ operations via matrix-vector multiplication. Thus, the per iteration cost of computing $\boldsymbol{u}_t$ given a factorized $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2\phi(1-\phi)^{-1}\boldsymbol{I}_d)$ is $\mathcal{O}(nd + d^2)$ arithmetic operations. Because there are at most n iterations, this leads to a total cost of $\mathcal{O}(n^2 d + nd^2)$ arithmetic operations. We remark that $\mathcal{O}(n)$ memory is required for storing vectors such as $\boldsymbol{u}_t(\mathcal{A}_t \setminus p_t)$.

Thus, an assignment may be sampled from the Gram–Schmidt Walk design using $\mathcal{O}(n^2 d)$ arithmetic computations and $\mathcal{O}(n + d^2)$ extra storage when implemented with these matrix factorizations. There are several practical considerations when implementing this algorithm. First, for what values of n and d is this practically feasible? Of course, this depends on the computing infrastructure which is available to experimenters, but roughly speaking, sampling from the Gram–Schmidt Walk is as computationally intensive as computing all pairs of inner products of covariates $\boldsymbol{x}_1, \boldsymbol{x}_2 \dots \boldsymbol{x}_n \in \mathbb{R}^d$. Computing these inner products requires $\mathcal{O}(n^2 d)$ arithmetic operations and computing this matrix of inner products $\boldsymbol{X}\boldsymbol{X}^\intercal$ is a pre-processing step of our implementation. The analysis above shows that the remainder of the algorithm requires roughly the same number of arithmetic operations. Thus, sampling from the Gram–Schmidt Walk should be practically feasible in cases where computing all inner products is practically feasible. A second practical consideration are the computational speed-ups for sampling more than one assignment from the design. When sampling many assignments from the Gram–Schmidt Walk, we may greatly reduce the run time by computing the initial cholesky factorization of $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2 \phi (1-\phi)^{-1} \boldsymbol{I}_d)$ and re-using it for each sample. Finally, we remark that although our focus is to speed up the Gram–Schmidt Walk when we use the augmented covariate vectors, similar matrix factorizations may also be used to decrease the asymptotic run time of the general Gram–Schmidt Walk.

A.2.5 Proof of asymptotic runtime (Proposition 2.11)

Proposition 2.11. *Assignments from the Gram–Schmidt Walk design can be sampled using $\mathcal{O}(n^2 d)$ arithmetic operations and $\mathcal{O}(n + d^2)$ additional storage.*

Proof. As detailed in Section A.2, these computational resource guarantees may be achieved by storing and maintaining a Cholesky factorization of the matrix $(\boldsymbol{X}_t^\intercal \boldsymbol{X}_t + \xi^2 \phi (1-\phi)^{-1} \boldsymbol{I}_d)$, where $\boldsymbol{X}_t$ denotes the row-submatrix of $\boldsymbol{X}$ with rows $\mathcal{A}_t \setminus p_t$. Constructing the matrix $\boldsymbol{X}^\intercal \boldsymbol{X}$ requires $\mathcal{O}(nd^2)$ arithmetic operations and $\mathcal{O}(d^2)$ space. Initially computing a Cholesky factorization of this matrix requires $\mathcal{O}(d^3)$ arithmetic operations and may be done in place. Updating the Cholesky factorization may be done using $\mathcal{O}(nd)$ arithmetic operations in place and this is done at most n times. Thus, constructing and maintaining the Cholesky factorization requires at most $\mathcal{O}(n^2 d)$ arithmetic operations and $\mathcal{O}(d^2)$ space, assuming that $d \leq n$.

Finally, computing the step direction $\boldsymbol{u}_t$ at each iteration requires $\mathcal{O}(nd)$ arithmetic operations and $\mathcal{O}(n)$ space given the above Cholesky factorization. This happens for at most n iterations, yielding a total of $\mathcal{O}(n^2 d)$ arithmetic operations and $\mathcal{O}(n)$ space. Thus, combining the computational requirements of maintaining the Cholesky factorizaiton and computing the step directions $\boldsymbol{u}_t$ yields a total requirement of $\mathcal{O}(n^2 d)$ arithmetic operations and $\mathcal{O}(n + d^2)$ additional storage to generate one assignment vector using the Gram–Schmidt Walk. □

A.3 Additional Proofs

In this section, we provide additional proofs of results in Chapter 2.

A.3.1 Analysis of the mean squared error (Theorem 2.14)

We begin by analyzing the mean squared error of the Horvitz–Thompson estimator under the GSW-Design. We start by presenting the relationship between the quadratic form in matrix $\boldsymbol{Q}$ and the loss of ridge regression.

Lemma A.8. *Let $\boldsymbol{X}$ be an arbitrary n-by-d matrix with maximum row norm $\xi = \max_{i\in[n]}\|\boldsymbol{x}_i\|$. For all $\phi \in (0,1)$ and $\boldsymbol{\mu} \in \mathbb{R}^n$,*

$$nL = \boldsymbol{\mu}^\intercal \boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{\mu}^\intercal\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}^\intercal\boldsymbol{X}\big)^{-1}\boldsymbol{\mu} = \min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\frac{1}{\phi}\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\xi^2}{1-\phi}\|\boldsymbol{\beta}\|^2\right].$$

Proof. Let $\boldsymbol{\beta}^*$ be the optimal linear function in the minimization term above. Note that multiplying the objective function by $\phi > 0$ does not change the minimizer $\boldsymbol{\beta}^*$, and so

$$\boldsymbol{\beta}^* = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\frac{1}{\phi}\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\xi^2}{1-\phi}\|\boldsymbol{\beta}\|^2\right] = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\xi^2\phi}{1-\phi}\|\boldsymbol{\beta}\|^2\right] ,$$

which has closed-form solution (see, e.g., Hastie et al., 2009, p. 64):

$$\boldsymbol{\beta}^* = \left(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\xi^2\phi}{1-\phi}\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{\mu} = \boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\boldsymbol{\mu} ,$$

where we have defined $\boldsymbol{R} = \boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\xi^2\phi}{1-\phi}\boldsymbol{I}$. We next consider each of the terms in the objective function when we substitute the optimal $\boldsymbol{\beta}^*$. The second term becomes

$$\frac{\xi^2}{1-\phi}\|\boldsymbol{\beta}^*\|^2 = \frac{\xi^2}{1-\phi}\big\|\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\boldsymbol{\mu}\big\|^2 = \frac{\xi^2}{1-\phi}\boldsymbol{\mu}^\intercal\boldsymbol{X}\boldsymbol{R}^{-2}\boldsymbol{X}^\intercal\boldsymbol{\mu}.$$

The first term becomes

$$\begin{aligned}
\frac{1}{\phi}\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}^*\|^2 &= \frac{1}{\phi}\big\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\boldsymbol{\mu}\big\|^2 = \frac{1}{\phi}\big\|\big(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\big)\boldsymbol{\mu}\big\|^2 \\
&= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\big(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\big)^2\boldsymbol{\mu} \\
&= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\big(\boldsymbol{I}-2\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal+\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\big)\boldsymbol{\mu} \\
&= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\big(\boldsymbol{I}-\boldsymbol{X}\big[2\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{R}^{-1}\big]\boldsymbol{X}^\intercal\big)\boldsymbol{\mu} \\
&= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\big(\boldsymbol{I}-\boldsymbol{X}\big[2\boldsymbol{R}^{-1}-\boldsymbol{R}^{-2}\boldsymbol{X}^\intercal\boldsymbol{X}\big]\boldsymbol{X}^\intercal\big)\boldsymbol{\mu} ,
\end{aligned}$$

where the last line follows from the fact that $\boldsymbol{R}^{-1}$ and $\boldsymbol{X}^\intercal\boldsymbol{X}$ commute. To see that the matrices $\boldsymbol{R}^{-1}$ and $\boldsymbol{X}\boldsymbol{X}^\intercal$ commute, first observe that $\boldsymbol{R} = \frac{\xi^2\phi}{1-\phi}\boldsymbol{I} + \boldsymbol{X}^\intercal\boldsymbol{X}$ has the same eigenvectors as $\boldsymbol{X}^\intercal\boldsymbol{X}$. It follows that $\boldsymbol{R}^{-1}$ also has the same eigenvectors as $\boldsymbol{X}^\intercal\boldsymbol{X}$. Thus, the two matrices $\boldsymbol{R}^{-1}$ and $\boldsymbol{X}^\intercal\boldsymbol{X}$ are simultaneously diagonalizable and therefore commute.

Substituting these separate calculations into the objective function, we obtain the optimal value

$$
\begin{aligned}
&\frac{1}{\phi}\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}^*\|^2 + \frac{\xi^2}{1-\phi}\|\boldsymbol{\beta}^*\|^2 \\
&\quad= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\big(\boldsymbol{I} - \boldsymbol{X}\big[2\boldsymbol{R}^{-1} - \boldsymbol{R}^{-2}\boldsymbol{X}^\intercal\boldsymbol{X}\big]\boldsymbol{X}^\intercal\big)\boldsymbol{\mu} + \frac{\xi^2}{1-\phi}\boldsymbol{\mu}^\intercal\boldsymbol{X}\boldsymbol{R}^{-2}\boldsymbol{X}^\intercal\boldsymbol{\mu} \\
&\quad= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\left(\boldsymbol{I} - \boldsymbol{X}\left[2\boldsymbol{R}^{-1} - \boldsymbol{R}^{-2}\boldsymbol{X}^\intercal\boldsymbol{X} - \frac{\phi\xi^2}{1-\phi}\boldsymbol{R}^{-2}\right]\boldsymbol{X}^\intercal\right)\boldsymbol{\mu} \\
&\quad= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\left(\boldsymbol{I} - \boldsymbol{X}\left[2\boldsymbol{R}^{-1} - \boldsymbol{R}^{-2}\left(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}\right)\right]\boldsymbol{X}^\intercal\right)\boldsymbol{\mu} \\
&\quad= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\big(\boldsymbol{I} - \boldsymbol{X}\big[2\boldsymbol{R}^{-1} - \boldsymbol{R}^{-2}\boldsymbol{R}\big]\boldsymbol{X}^\intercal\big)\boldsymbol{\mu} \\
&\quad= \frac{1}{\phi}\boldsymbol{\mu}^\intercal\big(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\big)\boldsymbol{\mu}
\end{aligned}
$$

To complete the proof, we apply the Woodbury identity which asserts that for appropriately sized matrices $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{C}$, $(\boldsymbol{I} + \boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{I} - \boldsymbol{U}\big(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{U}\big)^{-1}\boldsymbol{V}$, given that the inverses exist. Applying the Woodbury identity with $\boldsymbol{U} = \boldsymbol{X}$, $\boldsymbol{V} = \boldsymbol{X}^\intercal$, and $\boldsymbol{C} = \frac{1-\phi}{\xi^2\phi}\boldsymbol{I}$, we obtain

$$
\begin{aligned}
\frac{1}{\phi}\big(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\intercal\big) &= \frac{1}{\phi}\left(\boldsymbol{I} - \boldsymbol{X}\left(\frac{\xi^2\phi}{1-\phi}\boldsymbol{I} + \boldsymbol{X}^\intercal\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\intercal\right) \\
&= \frac{1}{\phi}\left(\boldsymbol{I} + \frac{\xi^{-2}(1-\phi)}{\phi}\boldsymbol{X}^\intercal\boldsymbol{X}\right)^{-1} = \big(\phi\boldsymbol{I} + \xi^{-2}(1-\phi)\boldsymbol{X}^\intercal\boldsymbol{X}\big)^{-1}.\square
\end{aligned}
$$

Using this lemma, we are now ready to establish the improved mean squared error analysis of the Horvitz–Thompson estimator under the GSW-Design.

Theorem 2.14. *The mean squared error under the* GSW-Design *is at most the minimum of the loss function of an implicit ridge regression of the sum of the potential outcome vectors* $\boldsymbol{\mu} = (\boldsymbol{a} + \boldsymbol{b})$ *on the covariates:*

$$
\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] \leq \frac{L}{n} \qquad \textit{where} \qquad L = \min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\frac{1}{\phi n}\big\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2 + \frac{\xi^2}{(1-\phi)n}\big\|\boldsymbol{\beta}\big\|^2\right].
$$

Proof. In Lemma 2.2, we established that the mean squared error of the Horvitz–Thompson estimator is a quadratic form in the covariance matrix of assignments,

$\mathrm{Cov}(\boldsymbol{z})$. We can obtain a bound on this matrix using the inequality in Theorem 2.12. The upper left n-by-n block of $\mathrm{Cov}(\boldsymbol{B}\boldsymbol{z})$ is $\phi\,\mathrm{Cov}(\boldsymbol{z})$. The corresponding block of the projection matrix $\boldsymbol{P}$ in Theorem 2.12 is $\phi\boldsymbol{Q}$ where

$$\boldsymbol{Q} = \left(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}.$$

If $\boldsymbol{A} \preceq \boldsymbol{B}$, then any two principal submatrices corresponding to the same row and column set S satisfy the inequality $\boldsymbol{A}_S \preceq \boldsymbol{B}_S$. It follows that $\mathrm{Cov}(\boldsymbol{z}) \preceq \boldsymbol{Q}$. Using the definition of the Loewner partial order together with LemmaA.8, we obtain

$$\mathbb{E}[(\tau - \widehat{\tau})^2] = \frac{1}{n^2}\boldsymbol{\mu}^\mathsf{T}\,\mathrm{Cov}(\boldsymbol{z})\boldsymbol{\mu} \leq \frac{1}{n^2}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{Q}\boldsymbol{\mu} = L/n \ . \qquad \square$$

A.3.2 Choosing the design parameter

In this section, we prove the results presented in Section 2.4.2, which illustrate how to choose the design parameter. Throughout this section, we let

$$\mathcal{L}(\phi) = \min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\frac{1}{\phi n}\left\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 + \frac{\xi^2}{(1-\phi)n}\left\|\boldsymbol{\beta}\right\|^2\right],$$

be the optimal ridge loss given design parameter ϕ. Similarly, we write

$$\mathcal{L}(\phi, \boldsymbol{\beta}) = \frac{1}{\phi n}\left\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 + \frac{\xi^2}{(1-\phi)n}\left\|\boldsymbol{\beta}\right\|^2$$

to refer to the ridge loss for a fixed design parameter ϕ and linear function $\boldsymbol{\beta}$.

The first result describes conditions under which lower mean squared error is achieved by setting $\phi < 1$.

Corollary 2.15. *If the scaled sum of cross-moments between covariates and potential outcomes is greater than the second moment of potential outcomes, $\xi^{-2}\|\boldsymbol{X}^\mathsf{T}\boldsymbol{\mu}\|^2 > \|\boldsymbol{\mu}\|^2$, then the design parameter ϕ that minimizes the mean squared error is less than one.*

Proof. We begin by letting

$$\boldsymbol{Q}(\phi) = \left(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right)^{-1}.$$

We can write $\mathcal{L}(\phi) = n^{-1}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{Q}(\phi)\boldsymbol{\mu}$, and

$$\frac{d\mathcal{L}(\phi)}{d\phi} = \frac{1}{n}\boldsymbol{\mu}^\mathsf{T}\boldsymbol{Q}(\phi)\left(\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T} - \boldsymbol{I}\right)\boldsymbol{Q}(\phi)\boldsymbol{\mu}.$$

Note that $\boldsymbol{Q}(1) = \boldsymbol{I}$, implying that

$$\frac{d\mathcal{L}(\phi)}{d\phi}\Big|_{\phi=1} > 0 \iff \boldsymbol{\mu}^\mathsf{T}\left(\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\mathsf{T} - \boldsymbol{I}\right)\boldsymbol{\mu} > 0 \iff \xi^{-2}\|\boldsymbol{X}^\mathsf{T}\boldsymbol{\mu}\|^2 > \|\boldsymbol{\mu}\|^2.$$

Note that $\mathcal{L}(1) = \|\boldsymbol{\mu}\|^2$, meaning that the inequality in Theorem 2.14 is an equality when $\phi = 1$. Thus, the derivative of the mean squared error coincide of the derivative of the bound at $\phi = 1$. □

The second result derives the asymptotic mean squared error for a fixed design parameter $\phi < 1$.

Corollary 2.16. *Let $\boldsymbol{\beta}_{\text{LS}} \in \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|$ be the best least squares linear approximator of the potential outcomes with smallest norm, and let $\boldsymbol{\varepsilon} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}$ be the errors of those approximations. Fix a design parameter $\phi < 1$. If $\|\boldsymbol{\beta}_{\text{LS}}\|^2 = o\big(\xi^{-2}n\big)$, then the normalized mean squared error under the* GSW-Design *is asymptotically upper bounded by*

$$\limsup_{n\to\infty} \left[n\,\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] - \frac{1}{\phi n}\|\boldsymbol{\varepsilon}\|^2 \right] \leq 0.$$

Let $\boldsymbol{\beta}_{\text{LS}} \in \arg\min_{\boldsymbol{\beta}} \big\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\big\|^2$. In the case $\boldsymbol{\beta}_{\text{LS}}$ is not uniquely defined, pick the solution of minimum norm. That is, $\boldsymbol{\beta}_{\text{LS}} = \boldsymbol{X}^{\dagger}\boldsymbol{\mu}$, where $\boldsymbol{X}^{\dagger}$ is the pseudoinverse of $\boldsymbol{X}$.

Proof. Using the mean squared error bound of Theorem 2.14 together with the definition of the ridge loss, we have that

$$n\,\mathbb{E}[(\tau - \widehat{\tau})^2] \leq \mathcal{L}(\phi) \leq \mathcal{L}(\phi, \boldsymbol{\beta}_{\text{LS}}) = \frac{1}{\phi n}\big\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}\big\|^2 + \frac{\xi^2}{(1-\phi)n}\big\|\boldsymbol{\beta}_{\text{LS}}\big\|^2 \ .$$

Using the definition of $\boldsymbol{\varepsilon} = \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}$ and rearranging terms yields

$$\left[n\,\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] - \frac{1}{\phi n}\|\boldsymbol{\varepsilon}\|^2 \right] \leq \frac{\xi^2}{(1-\phi)n}\big\|\boldsymbol{\beta}_{\text{LS}}\big\|^2.$$

The result is obtained by observing that for fixed $\phi > 0$ and $\|\boldsymbol{\beta}_{\text{LS}}\|^2 = o\big(\xi^{-2}n\big)$,

$$\lim_{n\to\infty} \frac{\xi^2}{(1-\phi)n}\big\|\boldsymbol{\beta}_{\text{LS}}\big\|^2 = 0.$$ □

Corollary 2.17. *Under the conditions of Corollary 2.16, the normalized mean squared error under the* GSW-Design *with the adaptive parameter choice of $\phi = \big(1 + \xi\|\boldsymbol{\beta}_{\text{LS}}\|/\|\boldsymbol{\varepsilon}\|\big)^{-1}$ is asymptotically upper bounded by*

$$\limsup_{n\to\infty} \left[n\,\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 \right] \leq 0.$$

Proof. Recall that $\boldsymbol{\beta}_{\text{LS}}$ is the vector in $\arg\min_{\boldsymbol{\beta}}\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|$ with smallest norm. Consider the design parameter $\phi^*_{\text{LS}} = \arg\min_{\phi} \mathcal{L}(\phi, \boldsymbol{\beta}_{\text{LS}})$ that minimizes the ridge loss at

$\boldsymbol{\beta}_{\text{LS}}$. This optimal design parameter satisfies the following first order condition:

$$\frac{d\mathcal{L}(\phi, \boldsymbol{\beta}_{\text{LS}})}{d\phi} = \frac{\xi^2}{(1-\phi)^2 n}\big\|\boldsymbol{\beta}_{\text{LS}}\big\|^2 - \frac{1}{\phi^2 n}\big\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}\big\|^2 = 0.$$

In particular, solving the first order condition yields the design parameter

$$\phi^*_{\text{LS}} = \frac{\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}\|}{\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}_{\text{LS}}\| + \xi\|\boldsymbol{\beta}_{\text{LS}}\|} = \frac{\|\boldsymbol{\varepsilon}\|}{\|\boldsymbol{\varepsilon}\| + \xi\|\boldsymbol{\beta}_{\text{LS}}\|} = \left(1 + \frac{\xi\|\boldsymbol{\beta}_{\text{LS}}\|}{\|\boldsymbol{\varepsilon}\|}\right)^{-1},$$

which is the parameter sequence specified in the proposition. Using the mean squared error bound of Theorem 2.14 together with the definition of the ridge loss, we have that

$$n\,\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] \leq \mathcal{L}(\phi^*_{\text{LS}}) \leq \mathcal{L}(\phi^*_{\text{LS}}, \boldsymbol{\beta}_{\text{LS}}) = \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 + \frac{2\xi}{n}\|\boldsymbol{\varepsilon}\| \times \|\boldsymbol{\beta}_{\text{LS}}\| + \frac{\xi^2}{n}\big\|\boldsymbol{\beta}_{\text{LS}}\big\|^2,$$

and rearranging terms yields

$$\Big[n\,\mathbb{E}\big[(\widehat{\tau} - \tau)^2\big] - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\Big] \leq \frac{2\xi}{n}\|\boldsymbol{\varepsilon}\| \times \|\boldsymbol{\beta}_{\text{LS}}\| + \frac{\xi^2}{n}\big\|\boldsymbol{\beta}_{\text{LS}}\big\|^2.$$

By assumption, we have that the linear coefficients are bounded as $\|\boldsymbol{\beta}_{\text{LS}}\|^2 = o\big(\xi^{-2} n\big)$, so that

$$\lim_{n\to\infty} \frac{\xi^2}{n}\big\|\boldsymbol{\beta}_{\text{LS}}\big\|^2 = 0.$$

Furthermore, $0 \leq \|\boldsymbol{\varepsilon}\| \leq \|\boldsymbol{\mu}\|$ by construction, so if $\|\boldsymbol{\mu}\|^2 = \mathcal{O}(n)$, then $\|\boldsymbol{\varepsilon}\| = \mathcal{O}(\sqrt{n})$. We therefore know that $\|\boldsymbol{\varepsilon}\| \times \|\boldsymbol{\beta}_{\text{LS}}\| = o\big(\xi^{-1} n\big)$, so

$$\lim_{n\to\infty} \frac{2\xi}{n}\|\boldsymbol{\varepsilon}\| \times \|\boldsymbol{\beta}_{\text{LS}}\| = 0,$$

which establishes the claim. □

Finally, the following result shows that the conditions of Corollaries 2.16 and 2.17 are implied by more standard conditions.

Lemma A.9. *If the second moment of the potential outcomes $\|\boldsymbol{\mu}\|^2/n$ stays bounded, the condition $\|\boldsymbol{\beta}_{\text{LS}}\|^2 = o\big(\xi^{-2} n\big)$ is satisfied if the maximum row norm ξ is asymptotically dominated by the smallest, non-zero singular value of $\boldsymbol{X}$.*

Proof. Because $\boldsymbol{\beta}_{\text{LS}}$ is the vector in $\arg\min_{\boldsymbol{\beta}}\|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|$ with smallest norm, we have $\boldsymbol{\beta}_{\text{LS}} = \boldsymbol{X}^\dagger\boldsymbol{\mu}$, where $\boldsymbol{X}^\dagger$ is the pseudoinverse of $\boldsymbol{X}$. Note that $\|\boldsymbol{\beta}_{\text{LS}}\| = \|\boldsymbol{X}^\dagger\boldsymbol{\mu}\| \leq \|\boldsymbol{X}^\dagger\| \times \|\boldsymbol{\mu}\|$, where $\|\boldsymbol{X}^\dagger\|$ denotes the operator norm. Recall that the operator norm is the largest singular value. Note that the largest singular value of $\boldsymbol{X}^\dagger$ is the same as the inverse of the smallest, non-zero singular value of $\boldsymbol{X}$. Let $\sigma_{\min}$ denote this smallest, non-zero singular value. We thus have $\|\boldsymbol{\beta}_{\text{LS}}\|^2 \leq \|\boldsymbol{\mu}\|^2/\sigma^2_{\min}$, and the

condition is satisfied if $\|\boldsymbol{\mu}\|^2/\sigma_{\min}^2 = o\big(\xi^{-2}n\big)$. When the second moment of the potential outcomes is bounded, $\|\boldsymbol{\mu}\|^2/n = \mathcal{O}(1)$, this collapses to $\xi = o(\sigma_{\min})$. □

A.3.3 Analysis of covariate balancing (Proposition 2.18)

We now present the proofs for a more refined analysis of the covariance balancing properties of the GSW-Design. In particular, we prove Proposition 2.18, which derives an upper bound on $\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})$ in terms of the weighted harmonic mean of two matrices. This result allows for finer insights on covariate balance, as discussed in Section 2.5.

Proposition 2.18. *Under the* GSW-Design*, the covariance matrix of $\boldsymbol{X}^\intercal \boldsymbol{z}$ is bounded in the Loewner order by*

$$\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z}) \preceq \Big(\phi(\boldsymbol{X}^\intercal \boldsymbol{X})^\dagger + (1-\phi)(\xi^2 \boldsymbol{\Pi})^\dagger\Big)^\dagger,$$

where $\boldsymbol{\Pi}$ is the orthogonal projection onto the rows of the covariate matrix $\boldsymbol{X}$ and $\boldsymbol{A}^\dagger$ denotes the pseudo-inverse of $\boldsymbol{A}$.

Proof. The proof follows a similar structure as the proof of Theorem 2.14, in that we also here extract the principal submatrices from the matrix inequality in Theorem 2.12. The lower right d-by-d block of $\mathrm{Cov}(\boldsymbol{B}\boldsymbol{z})$ is $\xi^{-2}(1-\phi)\,\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z})$. The corresponding d-by-d block of the matrix bound $\boldsymbol{P} = \boldsymbol{B}(\boldsymbol{B}^\intercal \boldsymbol{B})^{-1}\boldsymbol{B}^\intercal$ is

$$\xi^{-2}(1-\phi)\boldsymbol{X}^\intercal \big(\phi \boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\big)^{-1}\boldsymbol{X}.$$

After rearranging terms, this yields the inequality

$$\mathrm{Cov}(\boldsymbol{X}^\intercal \boldsymbol{z}) \preceq \boldsymbol{X}^\intercal \big(\phi \boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\big)^{-1}\boldsymbol{X}.$$

To prove the current proposition, we will show that we may re-write this matrix upper bound as

$$\boldsymbol{X}^\intercal \big(\phi \boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\big)^{-1}\boldsymbol{X} = \Big(\phi(\boldsymbol{X}^\intercal \boldsymbol{X})^\dagger + (1-\phi)(\xi^2 \boldsymbol{\Pi})^\dagger\Big)^\dagger$$

We do so by reasoning about the singular value decomposition of the covariate matrix $\boldsymbol{X}$. To this end, let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal$ be the singular value decomposition. We only consider the case where $d \leq n$, as the case where $d > n$ follows in a similar manner. If $d \leq n$, then $\boldsymbol{U}$ is a n-by-n orthogonal matrix, $\boldsymbol{\Sigma}$ is an n-by-n diagonal matrix with non-negative diagonal entries, and $\boldsymbol{V}$ is a d-by-n matrix with orthogonal rows. Using

the singular value decomposition and orthogonality properties of $\boldsymbol{U}$, we have that

$$\begin{aligned}
&\boldsymbol{X}^\intercal\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\big)^{-1}\boldsymbol{X} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal\big)^{-1}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal && \text{(SVD)} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal\big(\phi\boldsymbol{U}\boldsymbol{U}^\intercal + (1-\phi)\xi^{-2}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal\big)^{-1}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal && (\boldsymbol{U}\boldsymbol{U}^\intercal = \boldsymbol{I}) \\
&= \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal\big(\boldsymbol{U}\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\big)\boldsymbol{U}^\intercal\big)^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal && \text{(distributing } \boldsymbol{U}) \\
&= \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal\boldsymbol{U}\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\big)^{-1}\boldsymbol{U}^\intercal\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal && \text{(inverse and } \boldsymbol{U}^{-1} = \boldsymbol{U}^\intercal) \\
&= \boldsymbol{V}\boldsymbol{\Sigma}\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\big)^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal && (\boldsymbol{U}\boldsymbol{U}^\intercal = \boldsymbol{I})
\end{aligned}$$

We can compute the pseudo-inverse of this matrix as

$$\begin{aligned}
\Big(\boldsymbol{X}^\intercal\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\big)^{-1}\boldsymbol{X}\Big)^\dagger &= \Big(\boldsymbol{V}\boldsymbol{\Sigma}\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\big)^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\Big)^\dagger \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^\dagger\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\big)\boldsymbol{\Sigma}^\dagger\boldsymbol{V}^\intercal \\
&= \phi\boldsymbol{V}(\boldsymbol{\Sigma}^\dagger)^2\boldsymbol{V}^\intercal + (1-\phi)\xi^{-2}\boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}\boldsymbol{V}^\intercal\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\dagger\boldsymbol{V}^\intercal \\
&= \phi\boldsymbol{V}(\boldsymbol{\Sigma}^\dagger)^2\boldsymbol{V}^\intercal + (1-\phi)\xi^{-2}(\boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}\boldsymbol{V}^\intercal)^2,
\end{aligned}$$

where the third equality follows from distributing the outer matrices. We analyze each term separately, beginning with the left term. Note that

$$\boldsymbol{X}^\intercal\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^\intercal\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\intercal = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\intercal$$

and so by the orthogonality of rows of $\boldsymbol{V}$, one can check that

$$(\boldsymbol{X}^\intercal\boldsymbol{X})^\dagger = \boldsymbol{V}(\boldsymbol{\Sigma}^2)^\dagger\boldsymbol{V}^\intercal = \boldsymbol{V}(\boldsymbol{\Sigma}^\dagger)^2\boldsymbol{V}^\intercal.$$

The matrix in the second term is equal to the orthogonal projection matrix onto the row span of $\boldsymbol{X}$. To see this, observe that $\boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}\boldsymbol{V}^\intercal$ is the sum of the outer products of the right singular vectors corresponding to positive singular values. Because these vectors form an orthonormal basis for the row span of $\boldsymbol{X}$, the sum of their outer products is the projection matrix $\boldsymbol{\Pi}$. As $\boldsymbol{\Pi}^2 = \boldsymbol{\Pi} = \boldsymbol{\Pi}^\dagger$,

$$(1-\phi)\xi^{-2}(\boldsymbol{V}\boldsymbol{\Sigma}^\dagger\boldsymbol{\Sigma}\boldsymbol{V}^\intercal)^2 = (1-\phi)\xi^{-2}\boldsymbol{\Pi}^2 = (1-\phi)\xi^{-2}\boldsymbol{\Pi}^\dagger = (1-\phi)(\xi^2\boldsymbol{\Pi})^\dagger.$$

Putting these two terms together, we arrive at

$$\Big(\boldsymbol{X}^\intercal\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\big)^{-1}\boldsymbol{X}\Big)^\dagger = \phi(\boldsymbol{X}^\intercal\boldsymbol{X})^\dagger + (1-\phi)(\xi^2\boldsymbol{\Pi})^\dagger.$$

The proof is completed by taking the pseudoinverse of both sides. □

A.3.4 Second-order assignment probabilities (Lemma 2.22)

In order to construct a conservative estimator for the ridge loss L/n, we must establish which pairs of potential outcomes are never observed. In this section, we prove Lemma 2.22, which establishes that the only pairs of unobserved outcomes under the GSW-Design are the two outcomes for each unit. At the end of the section, we demonstrate a different bound on the second order assignment probabilities which does not depend on the sample size.

In order to show that the second order assignment probabilities are nonzero, we analyze the fractional assignments at the end of the first iteration. The main point of our argument is that for any pair of units $i, j \in [n]$ and assignments $v_i, v_j \in \{\pm 1\}$, there exists a choice of first pivot p_1 and first step size δ_1 so that, conditioned upon this choice, the probability of setting $z_i = v_i$ and $z_i = v_j$ in later iterations is nonzero. Our proof technique requires that each unit has equal probability of being assigned either treatment, i.e., $\Pr(z_i = 1) = 1/2$ for all $i \in [n]$. Recall that this occurs by setting the initial fractional assignment vector as $\boldsymbol{z}_1 = \boldsymbol{0}$.

We begin by presenting a basic lemma which bounds the joint probability of two binary random variables in terms of their marginal probabilities.

Lemma A.10. *For any discrete random variables X and Y,*

$$\Pr(X = x, Y = y) \geq \Pr(X = x) - \Pr(Y \neq y).$$

Proof. Observe that by probability axioms,

$$\Pr(X = x, Y = y) = \Pr(X = x) - \Pr(X = x, Y \neq y) \geq \Pr(X = x) - \Pr(Y \neq y). \qquad \square$$

Next, we derive a unit's marginal probability of assignment conditional on the outcome of the first iteration.

Lemma A.11. *The conditional probability that unit i is assigned to treatment $v_i \in \{\pm 1\}$ given the random decisions of the algorithm in the first iteration is*

$$\Pr(z_i = v_i \mid p_1, \delta_1) = \frac{1}{2}\Big(1 + v_i \boldsymbol{z}_2(i)\Big),$$

where we recall that $\boldsymbol{z}_2$ depends on p_1 and δ_1.

Proof. For any ± 1 random variable X and realization $v \in \{\pm 1\}$, we have that $\Pr(X = v) = \frac{1}{2}(1 - v\,\mathbb{E}[X])$. Using this expression and the martingale property of the fractional

assignments (Lemma 2.7), we have that

$$\Pr(z_i = v_i \mid p_1, \delta_1) = \frac{1}{2}\Big(1 + v_i \,\mathbb{E}[\boldsymbol{z}_T(i) \mid p_1, \delta_1]\Big) = \frac{1}{2}\Big(1 + v_i \boldsymbol{z}_2(i)\Big). \qquad \square$$

To reason about the fractional assignment $\boldsymbol{z}_2$, we have to reason about the step direction vector $\boldsymbol{u}_1$. We now demonstrate how to derive a matrix which contains all possible realizations of $\boldsymbol{u}_1$ as its columns, up to scaling.

The step direction $\boldsymbol{u}_1$ is completely determined by the choice of pivot p_1. Because we are only considering the first iteration, we drop the subscript 1 for now and, instead, write $\boldsymbol{u}_p$ to denote the step direction when the unit p is chosen as the first pivot. We claim that the step direction is given by

$$\boldsymbol{u}_p = \frac{\boldsymbol{Q}(:,p)}{\boldsymbol{Q}(p,p)} \quad \text{where} \quad \boldsymbol{Q} = \big(\boldsymbol{B}^\intercal \boldsymbol{B}\big)^{-1} = \Big(\phi \boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\Big)^{-1}$$

and $\boldsymbol{Q}(:,i)$ denotes the ith column of $\boldsymbol{Q}$ and $\boldsymbol{Q}(i,j)$ denotes the entry in the ith row and jth column of $\boldsymbol{Q}$. To see this, recall that the first step direction is obtained by setting the pivot coordinate $u_p(p) = 1$ and choosing the remaining coordinates as minimizers of the least squares problem

$$u_p([n] \setminus p) = \operatorname*{arg\,min}_{u_i : i \neq p} \Big\| \boldsymbol{b}_p + \sum_{i \neq p} u_i \boldsymbol{b}_i \Big\|^2.$$

When the vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_n$ are linearly independent, the solution is unique and the matrix $(\boldsymbol{B}^\intercal \boldsymbol{B})^{-1}$ exists. Recall that the augmented covariate vectors used in the Gram–Schmidt Walk design are linearly independent by construction for design parameters $\phi > 0$. By first-order optimality conditions, the entire vector $\boldsymbol{u}_p$ should satisfy the property that the vector

$$\boldsymbol{B}\boldsymbol{u}_p = \boldsymbol{b}_p + \sum_{i \neq p} u_i \boldsymbol{b}_i$$

is orthogonal to all $\boldsymbol{b}_i$ with $i \neq p$. That is,

$$0 = \Big\langle \boldsymbol{b}_i, \boldsymbol{B}\boldsymbol{u}_p \Big\rangle = \Big\langle \boldsymbol{B}\boldsymbol{e}_i, \boldsymbol{B}\boldsymbol{u}_p \Big\rangle = \Big\langle \boldsymbol{B}^\intercal \boldsymbol{B}\boldsymbol{e}_i, \boldsymbol{u}_p \Big\rangle \qquad \text{for all } i \neq p.$$

The columns of $\boldsymbol{Q} = (\boldsymbol{B}^\intercal \boldsymbol{B})^{-1}$ satisfy this orthogonality property, as

$$\langle \boldsymbol{B}^\intercal \boldsymbol{B}\boldsymbol{e}_i, (\boldsymbol{B}^\intercal \boldsymbol{B})^{-1}\boldsymbol{e}_p \rangle = \boldsymbol{e}_i^\intercal \boldsymbol{B}^\intercal \boldsymbol{B}(\boldsymbol{B}^\intercal \boldsymbol{B})^{-1}\boldsymbol{e}_p = \boldsymbol{e}_i^\intercal \boldsymbol{I}\boldsymbol{e}_p = \mathbf{1}[i = p].$$

Thus, by dividing the pth column $\boldsymbol{Q}(:,p)$ by the $\boldsymbol{Q}(p,p)$ diagonal entry, the pth coordinate becomes one and we obtain the direction $\boldsymbol{u}_p$.

In order to understand the step direction in the first iteration, we will prove properties of the matrix $\boldsymbol{Q}$. Before doing so, we introduce the following technical

lemma.

Lemma A.12. *Let $\boldsymbol{A}$ be an n-by-n positive semidefinite matrix with diagonal entries at most* 1*. For any $\gamma > 0$, the matrix $\boldsymbol{M} = (\boldsymbol{A} + \gamma \boldsymbol{I})^{-1}$ satisfies*

$$\boldsymbol{M}(i,j)^2 \leq (1+\gamma)^{-2} \boldsymbol{M}(i,i)\boldsymbol{M}(j,j) \qquad \text{for all } i \neq j \in [n].$$

Proof. Let $S = \{i, j\}$ be a pair of indices and define $R = [n] \setminus S$ to be the remaining indices. We are interested in the principal submatrix $\boldsymbol{M}(S,S)$. By using the expression for the inverse of a block matrix, we may express this principal submatrix as

$$\begin{aligned}
\boldsymbol{M}(S,S) &= \Big(\boldsymbol{A} + \gamma \boldsymbol{I}\Big)^{-1}(S,S) && \text{(definition of } \boldsymbol{M}) \\
&= \Big(\boldsymbol{A}(S,S) + \gamma \boldsymbol{I}_S - \boldsymbol{A}(S,R)\Big(\boldsymbol{A}(R,R) + \gamma \boldsymbol{I}_R\Big)^{-1}\boldsymbol{A}(R,S)\Big)^{-1} && \text{(block matrix inverse)} \\
&= \Big(\boldsymbol{A}(S,S) - \boldsymbol{A}(S,R)\Big(\boldsymbol{A}(R,R) + \gamma \boldsymbol{I}_R\Big)^{-1}\boldsymbol{A}(R,S) + \gamma \boldsymbol{I}_S\Big)^{-1} && \text{(rearranging terms)} \\
&= \Big(\boldsymbol{B}_S + \gamma \boldsymbol{I}_S\Big)^{-1}, && \text{(defining } \boldsymbol{B}_S)
\end{aligned}$$

where the matrices $\boldsymbol{I}_S$ and $\boldsymbol{I}_R$ are identity matrices of the appropriate sizes.

We claim that $\boldsymbol{B}_S$ is positive semidefinite with diagonal entries at most one. The positive semidefinite property follows because $\boldsymbol{B}_S$ is the Schur complement of $\boldsymbol{A}(R,R) + \gamma \boldsymbol{I}_R$ onto the block S. The matrix $\boldsymbol{A}(R,R) + \gamma \boldsymbol{I}_R$ is positive semidefinite so that the matrix $\boldsymbol{A}(S,R)(\boldsymbol{A}(R,R) + \gamma \boldsymbol{I}_R)^{-1}\boldsymbol{A}(R,S)$ is positive semidefinite and thus has non-negative diagonals. The diagonal entries of $\boldsymbol{A}(S,S)$ are at most one by assumption and because the diagonal entries of $\boldsymbol{A}(S,R)(\boldsymbol{A}(R,R) + \gamma \boldsymbol{I}_R)^{-1}\boldsymbol{A}(R,S)$ are non-negative, the diagonal entries of $\boldsymbol{B}_S$ are at most one.

Thus, the 2-by-2 matrix $\boldsymbol{M}(S,S)^{-1}$ may be expressed as

$$\boldsymbol{M}(S,S)^{-1} = \boldsymbol{B}_S + \gamma \boldsymbol{I} = \begin{pmatrix} \alpha & \eta \\ \eta & \beta \end{pmatrix} + \gamma \boldsymbol{I} = \begin{pmatrix} \alpha + \gamma & \eta \\ \eta & \beta + \gamma \end{pmatrix},$$

where the inequalities $\eta^2 \leq \alpha\beta$ and $\alpha, \beta \leq 1$ follow because $\boldsymbol{B}$ is positive semidefinite with diagonals at most 1. For $\gamma > 0$ this matrix is invertible, so

$$\boldsymbol{M}(S,S) = \frac{1}{\det \boldsymbol{M}(S,S)^{-1}} \begin{pmatrix} \beta + \gamma & -\eta \\ -\eta & \alpha + \gamma \end{pmatrix}.$$

If $\eta = 0$ then $\boldsymbol{M}(i,j) = 0$ so the desired inequality holds. Otherwise, $\eta^2 > 0$ and

using the properties of $\boldsymbol{B}_S$, we have that

$$\frac{\boldsymbol{M}(i,i)\boldsymbol{M}(j,j)}{\boldsymbol{M}(i,j)^2} = \frac{(\beta+\gamma)(\alpha+\gamma)}{\eta^2} \geq \frac{(\beta+\gamma)(\alpha+\gamma)}{\alpha\beta} = \Big(1+\frac{\gamma}{\beta}\Big)\Big(1+\frac{\gamma}{\alpha}\Big) \geq (1+\gamma)^2.$$

Rearranging terms yields the desired inequality. □

We now derive properties of the matrix $\boldsymbol{Q}$ which allow us to further reason about the step direction in the first iteration.

Lemma A.13. *The n-by-n matrix $\boldsymbol{Q} = (\boldsymbol{B}^\intercal\boldsymbol{B})^{-1} = (\phi\boldsymbol{I}+(1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal)^{-1}$ satisfies the following properties for all pairs of units $i \neq j \in [n]$:*

1. *Diagonal entries are lower bounded by $\boldsymbol{Q}(i,i) \geq 1$.*
2. *Off-diagonal entry upper bounded by $|\boldsymbol{Q}(i,j)| \leq \frac{1-\phi}{\phi}$.*
3. *All 2-by-2 principal submatrices admit the bound $\boldsymbol{Q}(i,j)^2 \leq (1-\phi)^2\boldsymbol{Q}(i,i)\boldsymbol{Q}(j,j)$.*

Proof. To begin proving the statements of the theorem, we derive the entries of the matrix $\boldsymbol{Q}$. By rearranging terms and using the Woodbury identity,

$$\begin{aligned}
\boldsymbol{Q} &= \Big(\phi\boldsymbol{I}_n + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\Big)^{-1} \\
&= \phi^{-1}\left[\boldsymbol{I}_n + \frac{(1-\phi)}{\phi\xi^2}\boldsymbol{X}\boldsymbol{X}^\intercal\right]^{-1} && \text{(rearranging terms)} \\
&= \phi^{-1}\left[\boldsymbol{I}_n - \frac{1-\phi}{\phi\xi^2}\boldsymbol{X}\Big(\boldsymbol{I}_d + \frac{1-\phi}{\phi\xi^2}\boldsymbol{X}^\intercal\boldsymbol{X}\Big)^{-1}\boldsymbol{X}^\intercal\right] && \text{(Woodbury identity)} \\
&= \phi^{-1}\left[\boldsymbol{I}_n - \boldsymbol{X}\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1}\boldsymbol{X}^\intercal\right]. && \text{(rearranging terms)}
\end{aligned}$$

So the entries of the matrix $\boldsymbol{Q}$ may be computed directly as

$$\boldsymbol{Q}(i,j) = \boldsymbol{e}_i^\intercal\boldsymbol{Q}\boldsymbol{e}_j = \phi^{-1}\left(\mathbf{1}[i=j] - \boldsymbol{x}_i^\intercal\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1}\boldsymbol{x}_j\right).$$

We will now bound a relevant quadratic form. Note that for any unit i, we have the following matrix bound: $\boldsymbol{X}^\intercal\boldsymbol{X} = \sum_{j=1}^n \boldsymbol{x}_j\boldsymbol{x}_j^\intercal \succeq \boldsymbol{x}_i\boldsymbol{x}_i^\intercal$. This implies the matrix inequality

$$\left(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\right)^{-1} \preceq \left(\boldsymbol{x}_i\boldsymbol{x}_i^\intercal + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\right)^{-1} \qquad \text{for all } i \in [n].$$

Set $\alpha = \phi\xi^2/(1-\phi)$. Using the matrix bound above and the Sherman–Morrison formula, we may bound the quadratic form as

$$\begin{aligned}
\boldsymbol{x}_i^\intercal\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \alpha\boldsymbol{I}_d\Big)^{-1}\boldsymbol{x}_i &\leq \boldsymbol{x}_i^\intercal\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\intercal + \alpha\boldsymbol{I}_d\Big)^{-1}\boldsymbol{x}_i && \text{(matrix bound above)}\\
&= \boldsymbol{x}_i^\intercal\left(\alpha^{-1}\boldsymbol{I}_d - \frac{\alpha^{-2}\boldsymbol{x}_i\boldsymbol{x}_i^\intercal}{1+\alpha^{-1}\|\boldsymbol{x}_i\|^2}\right)\boldsymbol{x}_i && \text{(Sherman–Morrison)}\\
&= \left(\alpha^{-1}\|\boldsymbol{x}_i\|^2 - \frac{\alpha^{-2}\|\boldsymbol{x}_i\|^4}{1+\alpha^{-1}\|\boldsymbol{x}_i\|^2}\right) && \text{(distributing terms)}\\
&= \frac{\|\boldsymbol{x}_i\|^2}{\alpha + \|\boldsymbol{x}_i\|^2} && \text{(rearranging terms)}\\
&= \frac{\|\boldsymbol{x}_i\|^2}{\frac{\phi\xi^2}{1-\phi} + \|\boldsymbol{x}_i\|^2} = \frac{\|\boldsymbol{x}_i\|^2/\xi^2}{\frac{\phi}{1-\phi} + \|\boldsymbol{x}_i\|^2/\xi^2} && \text{(substituting } \alpha)\\
&\leq \frac{1}{\frac{\phi}{1-\phi}+1} = 1-\phi,
\end{aligned}$$

where the second inequality follows from the facts that $\|\boldsymbol{x}_i\| \leq \max_{k\in[n]}\|\boldsymbol{x}_k\| = \xi$ and that for all $a > 0$, the function $f_a(y) = \frac{y^2}{a+y^2}$ is increasing for $y \geq 0$.

We now demonstrate the lower bound on diagonal entries of the matrix $\boldsymbol{Q}$. Using the closed form expression for the entries derived above and the bound on the quadratic form, we have

$$\boldsymbol{Q}(i,i) = \phi^{-1}\left(1 - \boldsymbol{x}_i^\intercal\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1}\boldsymbol{x}_i\right) \geq \phi^{-1}(1-(1-\phi)) = \phi^{-1}\phi = 1.$$

Next, we demonstrate the upper bound on the magnitude of the off-diagonal entries. Using the closed form expression for these entries derived above, the Cauchy-Schwartz inequality, and the above bound on the quadratic form, we have

$$\begin{aligned}
\boldsymbol{Q}(i,j)^2 &= \phi^{-2}\Big(\boldsymbol{x}_i^\intercal\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1}\boldsymbol{x}_j\Big)^2\\
&= \phi^{-2}\Big\langle\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1/2}\boldsymbol{x}_i, \Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1/2}\boldsymbol{x}_j\Big\rangle^2\\
&\leq \phi^{-2}\Big\|\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1/2}\boldsymbol{x}_i\Big\|^2\Big\|\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1/2}\boldsymbol{x}_j\Big\|^2 && \text{(Cauchy-Schwartz)}\\
&= \phi^{-2}\left(\boldsymbol{x}_i^\intercal\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1}\boldsymbol{x}_i\right)\left(\boldsymbol{x}_j^\intercal\Big(\boldsymbol{X}^\intercal\boldsymbol{X} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I}_d\Big)^{-1}\boldsymbol{x}_j\right)\\
&\leq \phi^{-2}(1-\phi)^2 = \Big(\frac{1-\phi}{\phi}\Big)^2, && \text{(bound above)}
\end{aligned}$$

which establishes the upper bound on the off diagonal entries, $|\boldsymbol{Q}(i,j)| \le (1-\phi)/\phi$.

Finally, we demonstrate the bound on 2-by-2 principal submatrices. Define

$$\boldsymbol{M} = \Big(\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal + \frac{\phi}{1-\phi}\boldsymbol{I}\Big)^{-1}.$$

By rearranging terms, we have

$$\boldsymbol{Q} = \Big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal\Big)^{-1} = (1-\phi)^{-1}\Big(\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal + \frac{\phi}{1-\phi}\boldsymbol{I}\Big)^{-1} = (1-\phi)^{-1}\boldsymbol{M}.$$

As $\xi = \max_{i\in[n]}\|\boldsymbol{x}_i\|$ and the diagonal entries of $\boldsymbol{X}\boldsymbol{X}^\intercal$ are $\|\boldsymbol{x}_i\|^2$, the matrix $\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal$ is positive semidefinite with diagonal entries at most 1. Note that the entries of $\boldsymbol{Q}$ are the same as the entries of $\boldsymbol{M}$, up to a common factor. If you are reading this part of my dissertation, email me with the subject line "sushi dinner" and I will buy you a sushi dinner when we are in the same town. This offer is valid for 5 years and to the first 5 people. Anyways, we may apply Lemma A.12 with $\boldsymbol{A} = \xi^{-2}\boldsymbol{X}\boldsymbol{X}^\intercal$ and $\gamma = \frac{\phi}{1-\phi}$ to obtain the third inequality in the statement of the proposition:

$$\boldsymbol{Q}(i,j)^2 \le \Big(1 + \frac{\phi}{1-\phi}\Big)^{-2}\boldsymbol{Q}(i,i)\boldsymbol{Q}(j,j) = (1-\phi)^2\boldsymbol{Q}(i,i)\boldsymbol{Q}(j,j). \qquad \square$$

We now have the tools to prove the proposition of interest, namely that all pairwise second order assignment probabilities are nonzero.

Lemma 2.22. *The second-order assignment probabilities are bounded away from zero under the* GSW-Design *for all pairs of units and all treatments:*

$$\Pr\big((z_i, z_j) = \boldsymbol{v}\big) > \frac{1}{4n}\min\Big\{\phi, \frac{\phi^2}{1-\phi}\Big\} \quad \textit{for all} \quad i \ne j \quad \textit{and all} \quad \boldsymbol{v} \in \{\pm 1\}^2.$$

Proof. Let $i, j \in [n]$ be two arbitrary but distinct units such that $\boldsymbol{Q}(i,i) \ge \boldsymbol{Q}(j,j)$, which is without loss of generality because of symmetry. We begin by lower bounding the second-order assignment probability conditioned on the random decisions made

in the first iteration, namely the first pivot p_1 and the step size δ_1:

$$\begin{aligned}
&\Pr(z_i = v_i, z_j = v_j \mid p_1, \delta_1) \\
&\quad \geq \Pr(z_i = v_i \mid p_1, \delta_1) - \Pr(z_j \neq v_j \mid p_1, \delta_1) && \text{(Lemma A.10)} \\
&\quad = \frac{1}{2}\Big(1 + v_i \,\mathbb{E}[\boldsymbol{z}_2(i) \mid p_1, \delta_1]\Big) - \frac{1}{2}\Big(1 - v_i \,\mathbb{E}[\boldsymbol{z}_2(i) \mid p_1, \delta_1]\Big) && \text{(Lemma A.11)} \\
&\quad = \frac{1}{2}\Big(v_i \,\mathbb{E}[\boldsymbol{z}_2(i) \mid p_1, \delta_1] + v_j \,\mathbb{E}[\boldsymbol{z}_2(j) \mid p_1, \delta_1]\Big) && \text{(rearranging terms)} \\
&\quad = \frac{1}{2}\Big(v_i \delta_1 u_1(i) + v_j \delta_1 u_1(j)\Big) && \text{(update rules, } \boldsymbol{z}_1 = \mathbf{0}) \\
&\quad = \frac{1}{2}\delta_1\Big(v_i u_1(i) + v_j u_1(j)\Big). && \text{(rearranging terms)}
\end{aligned}$$

We continue by conditioning on the event that the first pivot is unit i, so that $p_1 = i$. Once the pivot is determined, the first step direction $\boldsymbol{u}_1$ has been determined. We claim that when i is chosen as the pivot, the step direction $\boldsymbol{u}_1$ satisfies the following properties:

1. $u_1(i) = 1$
2. $\max_{k \in [n]} |u_1(k)| \leq \max\{1, \frac{1-\phi}{\phi}\}$
3. $|u_1(j)| \leq 1 - \phi$

The first property follows directly from $p_1 = i$. The second property follows by considering two types of coordinates of $\boldsymbol{u}_1$. As we already noted, the pivot coordinate is $u_1(i) = 1$. We bound the magnitude of non-pivot coordinates $k \neq i$ by combining statements (1) and (2) of Lemma A.13,

$$\big|u_1(k)\big| = \big|\langle \boldsymbol{u}_1, \boldsymbol{e}_k \rangle\big| = \left|\left\langle \frac{\boldsymbol{Q}(:,i)}{\boldsymbol{Q}(i,i)}, \boldsymbol{e}_k \right\rangle\right| = \left|\frac{\boldsymbol{Q}(k,i)}{\boldsymbol{Q}(i,i)}\right| = \frac{|\boldsymbol{Q}(k,i)|}{\boldsymbol{Q}(i,i)} \leq |\boldsymbol{Q}(k,i)| \leq \frac{1-\phi}{\phi}.$$

Combining these two yields that $|u_1(k)| \leq \max\{1, \frac{1-\phi}{\phi}\}$ for all $k \in [n]$. The third property follows by the assumption that $\boldsymbol{Q}(i,i) \geq \boldsymbol{Q}(j,j)$ and the third part of Lemma A.13. Namely, that

$$u_1(j)^2 = \frac{\boldsymbol{Q}(i,j)^2}{\boldsymbol{Q}(i,i)^2} \leq \frac{\boldsymbol{Q}(i,j)^2}{\boldsymbol{Q}(i,i)\boldsymbol{Q}(j,j)} \leq (1-\phi)^2,$$

which demonstrates that $|u_1(j)| \leq 1 - \phi$, as desired.

Because the initial fractional assignment is $\boldsymbol{z}_1(i) = \mathbf{0}$, the first step size δ_1 is randomly chosen as

$$\delta_1 = \begin{cases} \delta_1^+ = \Big(\max_{k\in[n]}|u_1(k)|\Big)^{-1} & \text{with probability } 1/2 \\ \delta_1^- = \Big(\max_{k\in[n]}|u_1(k)|\Big)^{-1} & \text{with probability } 1/2 \end{cases}$$

Suppose that we further condition on the choice of step size so that $\delta_1 v_i \geq 0$. We refer to this choice of step size as $\delta_1^{v_i}$. Conditioning on this choice of step size and using the properties of the step direction $\boldsymbol{u}_1$ yields

$$\begin{aligned}
2\Pr(z_i = v_i, z_j = v_j \mid p_1, \delta_1) &= \delta_1\Big(v_i u_1(i) + v_j u_1(j)\Big) && \text{(from above)} \\
&= \delta_1\Big(v_i + v_j u_1(j)\Big) && \text{(property 1 of } \boldsymbol{u}_1) \\
&= \Big(\max_{k\in[n]}|u_1(k)|\Big)^{-1}\Big(1 + v_i v_j u_1(j)\Big) && \text{(choice of } \delta_1) \\
&\geq \Big(\max\Big\{1, \frac{1-\phi}{\phi}\Big\}\Big)^{-1}\Big(1 + v_i v_j u_1(j)\Big) && \text{(property 2 of } \boldsymbol{u}_1) \\
&= \min\Big\{1, \frac{\phi}{1-\phi}\Big\}\Big(1 + v_i v_j u_1(j)\Big) \\
&\geq \min\Big\{1, \frac{\phi}{1-\phi}\Big\}\Big(1 - |u_1(j)|\Big) && (v_i v_j \in \{\pm 1\}) \\
&\geq \min\Big\{1, \frac{\phi}{1-\phi}\Big\} \cdot \phi && \text{(property 3 of } \boldsymbol{u}_1) \\
&= \min\Big\{\phi, \frac{\phi^2}{1-\phi}\Big\}
\end{aligned}$$

Recall that the first pivot is chosen uniformly at random from the set of all n units, so that the probability unit i is chosen as pivot is $1/n$. In addition, the step size considered above is chosen with probability $1/2$. Thus, the probability of choosing the pivot to be i and the step size to be $\delta_1^{v_i}$ is $1/2n$. Using this and the above inequalities,

we have that

$$\begin{aligned}\Pr(z_i = v_i, z_j = v_j) &\geq \Pr(p_1 = i, \delta_1 = \delta_1^{v_i}) \cdot \Pr(z_i = v_i, z_j = v_j \mid p_1 = i, \delta_1 = \delta_1^{v_i}) \\ &\geq \frac{1}{2n} \cdot \frac{1}{2}\delta_1\Big(v_i u_1(i) + v_j u_1(j)\Big) \\ &\geq \frac{1}{4n}\min\Big\{\phi, \frac{\phi^2}{1-\phi}\Big\}. \end{aligned}$$ □

The lower bound in Lemma 2.22 holds for all pairs of treatment assignments and any covariate matrix. In this sense, Lemma 2.22 is a worst-case bound, and we conjecture that it is tight. However, we have observed that most of the second-order assignment probabilities are considerably closer to 1/4 that what the bound in Lemma 2.22 suggests. Note that 1/4 is the value of all second order assignment probabilities when the individual assignments are independent. We provide some theoretical justification for this observation in Lemma A.14, which bounds the absolute difference between 1/4 and all second order assignment probabilities. In particular, for design parameters in the range $\phi \in [0.8, 1]$, Lemma A.14 provides a lower bound on all second order assignment probabilities which is independent of the sample size n. We remark that the fact that the lower bound becomes vacuous for $\phi < 0.8$ is a consequence of the proof technique in Lemma A.14, and it is not a reflection of a property of the design itself.

Lemma A.14. *The second-order assignment probabilities under the Gram–Schmidt Walk design satisfy*

$$\big|\Pr\big((z_i, z_j) = \boldsymbol{v}\big) - 1/4\big| \leq \frac{1-\phi}{\phi} \quad \textit{for all} \quad i \neq j \quad \textit{and all} \quad \boldsymbol{v} \in \{\pm 1\}^2.$$

Proof. Let $i, j \in [n]$ be two arbitrary but distinct units. Consider a vector $\boldsymbol{\mu} = (\mu(1), \dots \mu(n))$ such that $\mu(k) = 0$ for all $k \notin \{i, j\}$ and

$$\mu(i) = \sqrt{1/2} \qquad \text{and} \qquad \mu(j) = \begin{cases} \sqrt{1/2} & \text{if } \operatorname{Cov}(z_i, z_j) \geq 0, \\ -\sqrt{1/2} & \text{if } \operatorname{Cov}(z_i, z_j) < 0. \end{cases}$$

Observe that this implies that $\|\boldsymbol{\mu}\| = 1$.

The value of the quadratic form in $\operatorname{Cov}(\boldsymbol{z})$ evaluated at vector $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu}^\intercal \operatorname{Cov}(\boldsymbol{z})\boldsymbol{\mu} = \mu(i)^2 + \mu(j)^2 + 2\mu(i)\mu(j)\operatorname{Cov}(z_i, z_j) = 1 + \big|\operatorname{Cov}(z_i, z_j)\big|,$$

because $2\mu(i)\mu(j)\operatorname{Cov}(z_i, z_j) = \big|\operatorname{Cov}(z_i, z_j)\big|$.

From Corollary 2.13, the largest eigenvalue of $\operatorname{Cov}(\boldsymbol{z})$ is at most $1/\phi$, so by the Courant–Fischer theorem,

$$1 + \big|\operatorname{Cov}(z_i, z_j)\big| = \boldsymbol{\mu}^\intercal \operatorname{Cov}(\boldsymbol{z})\boldsymbol{\mu} \leq \|\boldsymbol{\mu}\|^2 \cdot \max_{\|\boldsymbol{v}\|=1} \frac{\boldsymbol{v}^\intercal \operatorname{Cov}(\boldsymbol{z})\boldsymbol{v}}{\boldsymbol{v}^\intercal \boldsymbol{v}} \leq \|\boldsymbol{\mu}\|^2/\phi = 1/\phi.$$

Rearranging this inequality yields

$$\big|\mathrm{Cov}(z_i, z_j)\big| \leq \frac{1-\phi}{\phi}.$$

Recall that each unit is assigned to either treatment with equal probability so that $\mathbb{E}[z_i] = \mathbb{E}[z_j] = 0$, which implies that $\mathrm{Cov}(z_i, z_j) = \mathbb{E}[z_i z_j]$. Thus, we have that for any treatment assignments $\boldsymbol{v} \in \{\pm 1\}^2$,

$$\big|\Pr\big((z_i, z_j) = \boldsymbol{v}\big) - 1/4\big| = \big|\mathbb{E}[z_i z_j]\big| \leq \frac{1-\phi}{\phi}. \qquad \square$$

A.3.5 Analyzing the kernelized GSW-Design

In this section, we prove Theorem 2.26, which bounds the mean squared error of the Horvitz–Thompson estimator under the kernelized GSW-Design presented in Section 2.7. To this end, we present a Lemma which derives the loss of the implicit kernelized ridge regression.

Lemma A.15. *Let $\mathcal{X}$ be the space of covariates, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel on the covariates, and let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the associated RKHS. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n \in \mathcal{X}$ be given with $\xi^2 = \max_{i\in[n]} k(\boldsymbol{x}_i, \boldsymbol{x}_i)$. For any vector $\boldsymbol{\mu} \in \mathbb{R}^n$, and $\phi \in [0,1]$, we have that*

$$\boldsymbol{\mu}^\intercal \Big(\phi \boldsymbol{I} + \xi^{-2}(1-\phi)\boldsymbol{K}\Big)^{-1} \boldsymbol{\mu} = \min_{f\in\mathcal{H}} \frac{1}{\phi} \sum_{i=1}^{n} \big(\boldsymbol{\mu}(i) - f(\boldsymbol{x}_i)\big)^2 + \frac{\xi^2}{1-\phi}\|f\|_{\mathcal{H}}^2 \ .$$

Proof. By the Representor Theorem (Schölkopf et al., 2001), the minimizer of the kernel ridge regression f^* has the form

$$f^*(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^* k(\boldsymbol{x}, \boldsymbol{x}_i) \ ,$$

where $\alpha_1^*, \alpha_2^*, \ldots \alpha_n^* \in \mathbb{R}$. Thus, the n-length vector of evaluations may be written as

$$\begin{bmatrix} f(\boldsymbol{x}_1) \\ f(\boldsymbol{x}_2) \\ \vdots \\ f(\boldsymbol{x}_n) \end{bmatrix} = \boldsymbol{K}\boldsymbol{\alpha}^* \ ,$$

where $\boldsymbol{K}$ is the symmetric matrix of kernel evaluations $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \ldots \alpha_n^*)$ is the vector of coefficients. Additionally, using the two properties of RKHS (defini-

tion 2.25), we can write the RKHS norm of f^* as

$$\begin{aligned}\|f^*\|_{\mathcal{H}} = \langle f^*, f^*\rangle_{\mathcal{H}} &= \Big\langle \sum_{i=1}^n \alpha_i^* k(\boldsymbol{x}, \boldsymbol{x}_i), \sum_{i=1}^n \alpha_i^* k(\boldsymbol{x}, \boldsymbol{x}_i)\Big\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^n \sum_{j=1}^n \langle k(\boldsymbol{x}, \boldsymbol{x}_i), k(\boldsymbol{x}, \boldsymbol{x}_j)\rangle_{\mathcal{H}} \alpha_i^* \alpha_j^* \\ &= \sum_{i=1}^n \sum_{j=1}^n k(\boldsymbol{x}_i, \boldsymbol{x}_j) \alpha_i^* \alpha_j^* \\ &= (\boldsymbol{\alpha}^*)^\intercal \boldsymbol{K} \boldsymbol{\alpha}^*\end{aligned}$$

Thus, we may re-write the infinite dimensional kernel ridge loss as a finite dimensional linear loss:

$$\begin{aligned}\min_{f\in\mathcal{H}} \frac{1}{\phi}\sum_{i=1}^n \big(\boldsymbol{\mu}(i) - f(\boldsymbol{x}_i)\big)^2 + \frac{\xi^2}{1-\phi}\|f\|_{\mathcal{H}}^2 &= \min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \frac{1}{\phi}\|\boldsymbol{\mu} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 + \frac{\xi^2}{1-\phi}\boldsymbol{\alpha}^\intercal \boldsymbol{K}\boldsymbol{\alpha} \\ &= \frac{1}{\phi}\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\Big[\|\boldsymbol{\mu} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 + \lambda \boldsymbol{\alpha}^\intercal \boldsymbol{K}\boldsymbol{\alpha}\Big] \ ,\end{aligned}$$

where $\lambda = \phi\xi^2/(1-\phi)$. The first order optimality condition may be expressed as

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}}\Big[\|\boldsymbol{\mu} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 + \lambda\boldsymbol{\alpha}^\intercal\boldsymbol{K}\boldsymbol{\alpha}\Big] &= \nabla_{\boldsymbol{\alpha}}\Big[\|\boldsymbol{\mu}\|^2 - 2\langle\boldsymbol{\mu}, \boldsymbol{K}\boldsymbol{\alpha}\rangle + \lambda\boldsymbol{\alpha}^\intercal\big(\boldsymbol{K}^2 + \lambda\boldsymbol{K}\big)\boldsymbol{\alpha}\Big] \\ &= -2\boldsymbol{K}\boldsymbol{\mu} + 2\big(\boldsymbol{K}^2 + \lambda\boldsymbol{K}\big)\boldsymbol{\alpha} = 0\end{aligned}$$

so that the optimal solution is $\boldsymbol{\alpha}^* = (\boldsymbol{K} + \frac{\phi\xi^2}{1-\phi}\boldsymbol{I})^{-1}\boldsymbol{\mu}$. The squared norm of the residual may be computed as

$$\|\boldsymbol{\mu} - \boldsymbol{K}\boldsymbol{\alpha}^*\|^2 = \Big\|\boldsymbol{\mu} - \boldsymbol{K}\Big(\boldsymbol{K} + \lambda\boldsymbol{I}\Big)^{-1}\boldsymbol{\mu}\Big\|^2 = \boldsymbol{\mu}^\intercal\Big(\boldsymbol{I} - \boldsymbol{K}\Big(\boldsymbol{K} + \lambda\boldsymbol{I}\Big)^{-1}\Big)^2\boldsymbol{\mu} \ .$$

Likewise, the norm of the function may be computed as

$$\boldsymbol{\alpha}^\intercal\boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{\mu}^\intercal\Big(\boldsymbol{K} + \lambda\boldsymbol{I}\Big)^{-1}\boldsymbol{K}\Big(\boldsymbol{K} + \lambda\boldsymbol{I}\Big)^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}^\intercal\boldsymbol{K}\Big(\boldsymbol{K} + \lambda\boldsymbol{I}\Big)^{-2}\boldsymbol{\mu} \ ,$$

where we used the fact that these matrices commute because they have the same

eigenvectors. Thus, the loss may be calculated as

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\Big[\|\boldsymbol{\mu}-\boldsymbol{K}\boldsymbol{\alpha}\|^2+\lambda\boldsymbol{\alpha}^\intercal\boldsymbol{K}\boldsymbol{\alpha}\Big]\\
&\quad=\boldsymbol{\mu}^\intercal\Big[\Big(\boldsymbol{I}-\boldsymbol{K}\Big(\boldsymbol{K}+\lambda\boldsymbol{I}\Big)^{-1}\Big)^2+\lambda\boldsymbol{K}\Big(\boldsymbol{K}+\lambda\boldsymbol{I}\Big)^{-2}\Big]\boldsymbol{\mu}\\
&\quad=\boldsymbol{\mu}^\intercal\Big[\boldsymbol{I}-2\boldsymbol{K}\big(\boldsymbol{K}+\lambda\boldsymbol{I}\big)^{-1}+\boldsymbol{K}^2\Big(\boldsymbol{K}+\lambda\boldsymbol{I}\Big)^{-2}+\lambda\boldsymbol{K}\Big(\boldsymbol{K}+\lambda\boldsymbol{I}\Big)^{-2}\Big]\boldsymbol{\mu}\\
&\quad=\boldsymbol{\mu}^\intercal\Big[\boldsymbol{I}-2\boldsymbol{K}\big(\boldsymbol{K}+\lambda\boldsymbol{I}\big)^{-1}+\boldsymbol{K}\Big(\boldsymbol{K}+\lambda\boldsymbol{I}\Big)^{-2}\Big(\boldsymbol{K}+\lambda\boldsymbol{I}\Big)\Big]\boldsymbol{\mu}\\
&\quad=\boldsymbol{\mu}^\intercal\Big[\boldsymbol{I}-2\boldsymbol{K}\big(\boldsymbol{K}+\lambda\boldsymbol{I}\big)^{-1}+\boldsymbol{K}\Big(\boldsymbol{K}+\lambda\boldsymbol{I}\Big)^{-1}\Big]\boldsymbol{\mu}\\
&\quad=\boldsymbol{\mu}^\intercal\Big[\boldsymbol{I}-\boldsymbol{K}\big(\boldsymbol{K}+\lambda\boldsymbol{I}\big)^{-1}\Big]\boldsymbol{\mu}
\end{aligned}$$

Using that $\lambda=\phi\xi^2/(1-\phi)$ and incorporating the $1/\phi$ term, we have that

$$\begin{aligned}
&\frac{1}{\phi}\cdot\min_{\boldsymbol{\alpha}\in\mathbb{R}^n}\Big[\|\boldsymbol{\mu}-\boldsymbol{K}\boldsymbol{\alpha}\|^2+\lambda\boldsymbol{\alpha}^\intercal\boldsymbol{K}\boldsymbol{\alpha}\Big]\\
&\qquad\qquad=\frac{1}{\phi}\boldsymbol{\mu}^\intercal\Big[\boldsymbol{I}-\boldsymbol{K}\Big(\boldsymbol{K}+\frac{\xi^2\phi}{1-\phi}\boldsymbol{I}\Big)^{-1}\Big]\boldsymbol{\mu}\\
&\qquad\qquad=\frac{1}{\phi}\boldsymbol{\mu}^\intercal\Big[\boldsymbol{I}-\boldsymbol{K}^{1/2}\Big(\boldsymbol{K}+\frac{\xi^2\phi}{1-\phi}\boldsymbol{I}\Big)^{-1}\boldsymbol{K}^{1/2}\Big]\boldsymbol{\mu}\\
&\qquad\qquad=\frac{1}{\phi}\boldsymbol{\mu}^\intercal\Big[\boldsymbol{I}+\frac{\xi^2(1-\phi)}{\phi}\boldsymbol{K}\Big]^{-1}\boldsymbol{\mu}\\
&\qquad\qquad=\boldsymbol{\mu}^\intercal\Big(\phi\boldsymbol{I}+\xi^2(1-\phi)\boldsymbol{K}\Big)^{-1}\boldsymbol{\mu}\ ,
\end{aligned}$$

where the second to last inequality follows from the Woodbury matrix identity. □

We are now ready to prove the bound on the mean squared error.

Theorem 2.26. *Let $\mathcal{X}$ be the space of covariates, let $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ be a kernel on the covariates, and let $(\mathcal{H},\langle\cdot,\cdot\rangle_{\mathcal{H}})$ be the associated RKHS. The mean squared error of the Horvitz–Thompson estimator under the kernelized* GSW-Design *is at most the minimum of the loss function of an implicit kernel ridge regression of the sum of the potential outcomes on the covariates:*

$$\mathbb{E}\big[(\widehat{\tau}-\tau)^2\big]\le\frac{1}{n}\cdot\min_{f\in\mathcal{H}}\left[\frac{1}{\phi}\cdot\frac{1}{n}\sum_{i=1}^{n}\big((a_i+b_i)-f(\boldsymbol{x}_i)\big)^2+\frac{\xi^2}{(1-\phi)n}\big\|f\big\|_{\mathcal{H}}^2\right].$$

Proof. In Lemma 2.2, we established that the mean squared error of the Horvitz–Thompson estimator is a quadratic form in the covariance matrix of assignments, $\mathrm{Cov}(\boldsymbol{z})$. We can obtain a bound on this matrix using the inequality in Theorem 2.12. The upper left n-by-n block of $\mathrm{Cov}(\boldsymbol{B}\boldsymbol{z})$ is $\phi\,\mathrm{Cov}(\boldsymbol{z})$. Under the kernelized GSW-

DESIGN, the corresponding block of the projection matrix $\boldsymbol{P}$ in Theorem 2.12 is $\phi\boldsymbol{Q}$ where

$$\boldsymbol{Q} = \big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{K}\big)^{-1}.$$

If $\boldsymbol{A} \preceq \boldsymbol{B}$, then any two principal submatrices corresponding to the same row and column set S satisfy the inequality $\boldsymbol{A}_S \preceq \boldsymbol{B}_S$. It follows that $\mathrm{Cov}(\boldsymbol{z}) \preceq \boldsymbol{Q}$. Using the definition of the Loewner partial order together with LemmaA.8, we obtain

$$\mathbb{E}[(\tau - \widehat{\tau})^2] = \frac{1}{n^2}\boldsymbol{\mu}^\intercal \mathrm{Cov}(\boldsymbol{z})\boldsymbol{\mu} \leq \frac{1}{n^2}\boldsymbol{\mu}^\intercal\boldsymbol{Q}\boldsymbol{\mu} = \frac{1}{n^2}\boldsymbol{\mu}^\intercal\big(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{K}\big)^{-1}\boldsymbol{\mu}\ .$$

The proof is completed by using Lemma A.15, which shows that the right hand side is equal to the loss of the kernel ridge regression. □

Appendix B

Appendix for Optimized Variance Estimation under Interference and Complex Experimental Designs

B.1 Extension to General Linear Estimators

In this section, we demonstrate how to extend our analyses to general linear estimators. A general linear estimator may be written as

$$\widehat{\tau} = \sum_{i=1}^{n} \sum_{e \in \Delta} w_i \mathbf{1}[d_i(\boldsymbol{z}) = e] y_i(e).$$

As before, we introduce two variables to make this expression more manageable. For each unit-exposure pair $(i, e) \in [n] \times \Delta$, we define the variables

$$v_{i,e} = w_i \mathbf{1}[d_i(\boldsymbol{z}) = e] \qquad \text{and} \qquad \theta_{i,e} = y_i(e) \ .$$

Note that the treatment effect estimator may be written as

$$\widehat{\tau} = \sum_{i=1}^{n} \sum_{e \in \Delta} v_{i,e} \theta_{i,e} = \langle \boldsymbol{v}, \boldsymbol{\theta} \rangle \ ,$$

where $\boldsymbol{v}$ and $\boldsymbol{\theta}$ are vectors obtained by collected the variables $v_{i,e}$ and $\theta_{i,e}$, respectively. These vectors are K-dimensional, where $K = n \cdot |\Delta|$. There are a number of ways to order these coefficients into vectors, but that ordering itself doesn't matter. As before, the advantage of writing linear estimators in this way is that the potential outcomes are collected in the (deterministic) vector $\boldsymbol{\theta}$ and the randomness in the design and estimator is isolated to the random vector $\boldsymbol{v}$.

Using this notation, it is now simple to derive the variance of the linear estimator as a quadratic form, $\mathrm{Var}(\widehat{\tau}) = \boldsymbol{\theta}^{\mathsf{T}} \mathrm{Cov}(\boldsymbol{v}) \boldsymbol{\theta}$. The results that we have established in

Chapter 3 now follow, as we have established that the variance of a general linear estimator is a quadratic form in the potential outcome vector.

B.2 Additional Proofs

In this section, we prove the results presented in Chapter 3.

B.2.1 Design compatibility

The first result is that design-compatibility of a quadratic form characterizes when a unbiased estimator of the quadratic form exists.

Proposition 3.3. *An unbiased estimator exists for a quadratic form if and only if it is design compatible.*

Proof. If the quadratic form is design-compatible, then the Horvitz–Thompson estimator is unbiased, as shown in Section 3.5.

Suppose that a quadratic form is not design-compatible. For sake of contradiction, suppose that there exists an unbiased estimator $\widehat{\tau}$ so that $\mathbb{E}[\widehat{\tau}] = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{A}\boldsymbol{\theta}$. Let $i, j \in P$ be such that $\Pr(i, j \in S) = 0$. We know that such a pair exists because the quadratic from is design incompatible. Use the law of iterated expectation to write

$$\mathbb{E}[\widehat{\tau}] = \Pr(i \in S) f(\boldsymbol{\theta}) + \Pr(i \notin S) g(\boldsymbol{\theta}),$$

where $f(\boldsymbol{\theta}) = \mathbb{E}[\widehat{\tau} \mid i \in S]$ and $g(\boldsymbol{\theta}) = \mathbb{E}[\widehat{\tau} \mid i \notin S]$. We know that $g(\boldsymbol{\theta})$ does not depend on θ_i because the coordinate is never observed when $i \notin S$. Recall that $\Pr(i, j \in S) = 0$, so we know that $f(\boldsymbol{\theta})$ does not depend on θ_j because the coordinate is never observed when $i \in S$. It is not possible to write the quadratic form $\boldsymbol{\theta}^\mathsf{T}\boldsymbol{A}\boldsymbol{\theta}$ with $a_{ij} \neq 0$ as a linear combination of two functions where one does not depend on θ_i and the other does not depend on θ_j. □

B.2.2 Selection of variance bounds using OPT-VB

The following result gives conditions on the objective so that the variance bound returned by OPT-VB is admissible.

Theorem 3.8. *If g is strictly monotone, then* OPT-VB *returns a variance bound that is conservative, design compatible and admissible.*

Proof. By definition of the program, $\boldsymbol{S}^* \in \mathcal{S}$ and so the resulting variance bound $\boldsymbol{B}^*$ is conservative and design compatible. For sake of contradiction, assume that $\boldsymbol{B}^*$ is not admissible. Then, there exists a conservative and design compatible bound $\boldsymbol{B} \in \mathcal{B}$ with corresponding slack matrix $\boldsymbol{S} \in \mathcal{S}$ such that $\boldsymbol{B} = \boldsymbol{B}^* - \boldsymbol{Q}$ for some nonzero positive semidefinite matrix $\boldsymbol{Q}$. By subtracting $\boldsymbol{A}$ from both sides, we can

write this equality in terms of slack matrices, $\boldsymbol{S} = \boldsymbol{S}^* - \boldsymbol{Q}$. By assumption, the objective g is strictly monotone and $\boldsymbol{Q}$ is nonzero positive semidefinite so that

$$g(\boldsymbol{S}) < g(\boldsymbol{S} + \boldsymbol{Q}) = g(\boldsymbol{S}^*).$$

However, we have arrived at a contradiction, as $\boldsymbol{S}^*$ is a minimizer of g over $\mathcal{S}$. Thus, $\boldsymbol{B}^*$ is admissible. □

The following proposition demonstrates that all Schatten p-norms yield admissible variance bounds when used as objectives.

Proposition 3.9. *For all $p \in [1, \infty)$, the Schatten p-norm objective $g(\boldsymbol{S}) = \|\boldsymbol{A}+\boldsymbol{S}\|_p$ is strictly monotone, ensuring that the variance bound produced by* OPT-VB *using g is admissible.*

Proof. Let $\boldsymbol{A}$ be an n-by-n positive semidefinite matrix and let $\boldsymbol{S}$ be an n-by-n positive semidefinite matrix which is nonzero. Let the eigenvalues fo $\boldsymbol{A}+\boldsymbol{S}$ be denoted $\mu_1, \mu_2, \dots \mu_{2n}$ and let the eigenvalues of $\boldsymbol{A}$ be denoted $\lambda_1, \lambda_2, \dots \lambda_{2n}$. We show that for each $1 \leq \ell \leq 2n$, we have that $\mu_\ell \geq \lambda_\ell$ and at least one of the inequalities is strict.

Because $\boldsymbol{S}$ is positive semidefinite, it follows that $\mu_\ell \geq \lambda_\ell$ for each $1 \leq \ell \leq 2n$. We now show that at least one of these inequalities is strict. Recall that the trace of a matrix is the sum of the eigenvalues so that

$$\sum_{\ell=1}^{2n} \lambda_\ell = \operatorname{tr}(\boldsymbol{A}) < \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{S}) = \operatorname{tr}(\boldsymbol{A} + \boldsymbol{S}) = \sum_{\ell=1}^{2n} \mu_\ell \ ,$$

where the strict inequality follows from the fact that $\boldsymbol{S}$ is nonzero and positive semidefinite. Thus, the inequality is strict for at least one $\ell \leq 2n$.

The strict-monotonicity of $g(\boldsymbol{S})$ is now established using the result above and observing that the function $x \to x^p$ is strictly monotone on the real line. □

Finally, we demonstrate that a variance bound is admissible if and only if it may be obtained by running OPT-VB with a positive definite linear objective.

Theorem 3.10. *A variance bound $\boldsymbol{B}$ is admissible if and only if can be obtained from* OPT-VB *using the objective function $g(\boldsymbol{S}) = \langle \boldsymbol{S}, \boldsymbol{W} \rangle$ for some positive definite targeting matrix $\boldsymbol{W}$.*

Proof. To show that every variance bound obtained using the objective $g(\boldsymbol{S}) = \langle \boldsymbol{W}, \boldsymbol{S} \rangle$ is admissible, we show that g is strictly monotone and appeal to Theorem 3.8. Let $\boldsymbol{Q}$ be a nonzero positive semidefinite matrix. Then we may write

$$g(\boldsymbol{S} + \boldsymbol{Q}) - g(\boldsymbol{S}) = \langle \boldsymbol{W}, \boldsymbol{S} + \boldsymbol{Q} \rangle - \langle \boldsymbol{W}, \boldsymbol{S} \rangle = \langle \boldsymbol{W}, \boldsymbol{Q} \rangle,$$

where the last equality follows by linearity of the inner product. Let the eigendecomposition of $\boldsymbol{Q}$ be given as $\boldsymbol{Q} = \sum_{i=1}^{2n} \lambda_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^\intercal$. Then, the inner product may be

rewritten as

$$\langle \boldsymbol{W}, \boldsymbol{Q} \rangle = \langle \boldsymbol{W}, \sum_{i=1}^{2n} \lambda_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^\intercal \rangle = \sum_{i=1}^{2n} \lambda_i \boldsymbol{\eta}_i^\intercal \boldsymbol{W} \boldsymbol{\eta}_i .$$

Because $\boldsymbol{W}$ is positive definite, each of the $\boldsymbol{\eta}_i^\intercal \boldsymbol{W} \boldsymbol{\eta}_i$ terms are positive. Likewise, because $\boldsymbol{Q}$ is positive semidefinite and nonzero, there exists at least one positive eigenvalue $\lambda_i > 0$. This establishes that $g(\boldsymbol{S} + \boldsymbol{Q}) > g(\boldsymbol{S})$ so that g is strictly monotone. Thus, by Theorem 3.8, the resulting variance bound is admissible.

Now, let us suppose that $\boldsymbol{B}$ is an admissible variance bound and write the corresponding slack matrix as $\boldsymbol{S} = \boldsymbol{B} - \boldsymbol{A}$. Define the set

$$\mathcal{F}_{\boldsymbol{B}} = \left\{ \tilde{\boldsymbol{B}} = \boldsymbol{B} - \boldsymbol{Q} : \boldsymbol{Q} \text{ is nonzero and positive semidefinite} \right\} .$$

Because $\boldsymbol{B}$ is admissible, there does not exist another variance bound $\tilde{\boldsymbol{B}} \in \mathcal{B}$ which is in the set $\mathcal{F}_{\boldsymbol{B}}$. In other words, the intersection of $\mathcal{F}_{\boldsymbol{B}}$ and $\mathcal{B}$ is empty. Because the two sets $\mathcal{F}_{\boldsymbol{B}}$ and $\mathcal{B}$ are disjoint and convex, there exists a separating hyperplane between them. That is, there exists a matrix $\boldsymbol{W}$ and a scalar α so that

$$\begin{aligned} \langle \boldsymbol{W}, \boldsymbol{B} \rangle &\geq \alpha \text{ for all } \boldsymbol{B} \in \mathcal{B} \\ \langle \boldsymbol{W}, \tilde{\boldsymbol{B}} \rangle &< \alpha \text{ for all } \tilde{\boldsymbol{B}} \in \mathcal{F}_{\boldsymbol{B}} \end{aligned}$$

Let us first establish that $\langle \boldsymbol{W}, \boldsymbol{B} \rangle = \alpha$. For sake of contradiction, suppose that $\langle \boldsymbol{W}, \boldsymbol{B} \rangle = \alpha + \epsilon$ for some $\epsilon > 0$. Consider the matrix $\boldsymbol{H} = \boldsymbol{B} - \beta \cdot \boldsymbol{I}$, where $\beta = \frac{\epsilon}{2 \operatorname{tr}(\boldsymbol{W})}$. It follows that $\boldsymbol{H}$ is in the set $\mathcal{F}_{\boldsymbol{B}}$. However, we can compute

$$\langle \boldsymbol{W}, \boldsymbol{H} \rangle = \langle \boldsymbol{W}, \boldsymbol{B} - \beta \cdot \boldsymbol{I} \rangle = \langle \boldsymbol{W}, \boldsymbol{B} \rangle - \beta \langle \boldsymbol{W}, \boldsymbol{I} \rangle \geq \alpha + \epsilon/2 \ ,$$

which is a contradiction of the separating hyperplane. Thus, $\langle \boldsymbol{W}, \boldsymbol{B} \rangle = \alpha$.

Let us next establish that $\boldsymbol{W}$ is positive definite. Let the eigenvalue decomposition of $\boldsymbol{W}$ be given as $\boldsymbol{W} = \sum_{i=1}^{2n} \lambda_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^\intercal$. For sake of contradiction, suppose that one of the eigenvalues λ_k is non-positive. Let $\boldsymbol{\eta}_k$ be the corresponding eigenvector. Consider the matrix $\boldsymbol{H} = \boldsymbol{B} - \boldsymbol{\eta}_k \boldsymbol{\eta}_k^\intercal$, which is in the set $\mathcal{F}_{\boldsymbol{B}}$. However, we can obtain that

$$\langle \boldsymbol{W}, \boldsymbol{H} \rangle = \langle \boldsymbol{W}, \boldsymbol{B} - \boldsymbol{\eta}_k \boldsymbol{\eta}_k^\intercal \rangle = \langle \boldsymbol{W}, \boldsymbol{B} \rangle - \boldsymbol{\eta}_k^\intercal \boldsymbol{W} \boldsymbol{\eta}_k = \alpha - \lambda_k \geq \alpha \ ,$$

which is a contradiction of the separating hyperplane. Thus, $\boldsymbol{W}$ is positive definite.

Finally, we show that $\boldsymbol{B}$ may be obtained by using the objective $g(\boldsymbol{S}) = \langle \boldsymbol{W}, \boldsymbol{S} \rangle$. First, observe that the corresponding slack matrix takes value

$$g(\boldsymbol{S}) = \langle \boldsymbol{W}, \boldsymbol{S} \rangle = \langle \boldsymbol{W}, \boldsymbol{A} + \boldsymbol{S} \rangle - \langle \boldsymbol{W}, \boldsymbol{A} \rangle = \langle \boldsymbol{W}, \boldsymbol{B} \rangle - \langle \boldsymbol{W}, \boldsymbol{A} \rangle = \alpha - \langle \boldsymbol{W}, \boldsymbol{A} \rangle \ .$$

By the separating hyperplane, any other slack matrix $\tilde{\boldsymbol{S}} \in \mathcal{S}$ (with corresponding

bound matrix $\tilde{\boldsymbol{B}} \in \mathcal{B}$) has objective value at most

$$g(\tilde{\boldsymbol{S}}) = \langle \boldsymbol{W}, \tilde{\boldsymbol{S}} \rangle = \langle \boldsymbol{W}, \boldsymbol{A} + \tilde{\boldsymbol{S}} \rangle - \langle \boldsymbol{W}, \boldsymbol{A} \rangle = \langle \boldsymbol{W}, \tilde{\boldsymbol{B}} \rangle - \langle \boldsymbol{W}, \boldsymbol{A} \rangle \geq \alpha - \langle \boldsymbol{W}, \boldsymbol{A} \rangle \ .$$

Thus, $\boldsymbol{S}$ is a minimizer of g over $\mathcal{S}$. □

The following proposition guarantees that positive combinations of monotone functions is strictly monotone, provided that one of the functions is strictly monotone.

Proposition 3.11. *If a set of m functions $g_1, \ldots, g_m$ are monotone and at least one of the functions are strictly monotone, then for any set of positive coefficients $\gamma_1, \ldots, \gamma_m$, the function $g_c = \sum_{i=1}^m \gamma_i g_i$ is strictly monotone. Thus,* OPT-VB *returns a variance bound that is conservative, design compatible and admissible when called with the composite objective g_c.*

Proof. We begin by showing strict monotonicity of the function g_c. Let $\boldsymbol{A}$ be a positive semidefinite matrix and let $\boldsymbol{S}$ be a positive semidefinite and nonzero matrix. Without loss of generality, suppose that g_1 is the strictly monotone function. We have that

$$\begin{aligned} g_c(\boldsymbol{A}) &= \gamma_1 g_1(\boldsymbol{A}) + \sum_{i=2}^m \gamma_i g_i(\boldsymbol{A}) \\ &< \gamma_1 g_1(\boldsymbol{A} + \boldsymbol{S}) + \sum_{i=2}^m \gamma_i g_i(\boldsymbol{A}) \\ &\leq \gamma_1 g_1(\boldsymbol{A} + \boldsymbol{S}) + \sum_{i=2}^m \gamma_i g_i(\boldsymbol{A} + \boldsymbol{S}) \\ &= g_c(\boldsymbol{A} + \boldsymbol{S}) \ , \end{aligned}$$

where the equalities follow by definition of g_c, the strict inequality follows by strict monotonicity of g_1 together with positivity of γ_1 and the next inequality follows by monotonicity of g_i for $i = 2, \ldots m$ and the non-negativity of the coefficients. Thus, g_c is strictly monotone.

The properties of the variance bound obtained by using OPT-VB follow from Proposition 3.8. □

B.2.3 Consistent estimation of variance bounds

In the main body, Proposition 3.13 gave an upper bound on the mean squared error of the Horvitz–Thompson estimator of the variance bound. This upper bound was the product of three terms, corresponding to the design, the variance bound, and the potential outcomes. The bound depended on the largest magnitude of the potential outcomes. In this section, we demonstrate how the bound may be generalized so that only second moment conditions are required on the potential outcomes.

Before continuing, we introduce the entry-wise $L_{p,q}$ matrix norm. Given an $n \times n$ matrix $\boldsymbol{A}$, the $L_{p,q}$ *matrix norm* is defined as

$$\|\boldsymbol{A}\|_{p,q} = \Big(\sum_{j=1}^{n}\Big(\sum_{i=1}^{n}|a_{i,j}|^p\Big)^{q/p}\Big)^{1/q} \ .$$

When $p = q = 2$, then we recover the usual Frobenius norm. The more general finite sample bound on the MSE of the variance bound estimator appears below as Proposition B.1.

Proposition B.1. *Suppose that the variance bound $\boldsymbol{B}$ is design-compatible. Then, for any integers p and q with $1/p + 1/q = 1$, the mean squared error of the Horvitz–Thompson estimator may be bounded as*

$$\mathbb{E}[(\,VB(\boldsymbol{\theta}) - \widehat{VB}(\boldsymbol{\theta}))^2] \leq \|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\|_* \cdot \|\boldsymbol{B}\|_{2p,2}^2 \cdot \Big(\sum_{i=1}^{2n}\theta_i^{2q}\Big)^{2/q} \ .$$

Proof. Because $\Pr(i, j \in S) > 0$ for all pairs $(i.j) \in \bar{\Omega}$, the Horvitz–Thompson estimator is unbiased. Thus, the mean squared error is equal to the variance of the estimator, which may be computed as

$$\begin{aligned}
\mathbb{E}\Big[(\mathrm{VB}(\boldsymbol{\theta}) - \widehat{\mathrm{VB}}(\boldsymbol{\theta}))^2\Big] &= \mathrm{Var}(\widehat{\mathrm{VB}}(\boldsymbol{\theta})) \\
&= \mathrm{Var}\Big(\sum_{(i,j)\in\bar{\Omega}} \mathbf{1}[i,j \in S]\frac{b_{ij}\theta_i\theta_j}{\Pr(i,j\in S)}\Big) \\
&= \sum_{\substack{(i,j)\in\bar{\Omega}\\(k,\ell)\in\bar{\Omega}}} \mathrm{Cov}\Big(\mathbf{1}[i,j \in S]\frac{b_{ij}\theta_i\theta_j}{\Pr(i,j\in S)}, \mathbf{1}[k,\ell \in S]\frac{b_{k\ell}\theta_k\theta_\ell}{\Pr(k,\ell\in S)}\Big) \\
&= \sum_{\substack{(i,j)\in\bar{\Omega}\\(k,\ell)\in\bar{\Omega}}} \mathrm{Cov}\Big(\frac{\mathbf{1}[i,j \in S]}{\Pr(i,j\in S)}, \frac{\mathbf{1}[k,\ell \in S]}{\Pr(k,\ell\in S)}\Big)(b_{ij}\theta_i\theta_j)(b_{k\ell}\theta_k\theta_\ell) \\
&= b_{\bar{\Omega}}^{\intercal}\,\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})b_{\bar{\Omega}} \ ,
\end{aligned}$$

where $z_{\bar{\Omega}}$ and $b_{\bar{\Omega}}$ are vectors of length $|\bar{\Omega}|$, whose coordinates are indexed by pairs $(i,j) \in \bar{\Omega}$. The entries of $z_{\bar{\Omega}}$ are the inverse propensity weighted indicator vector $z_{\bar{\Omega}}(i,j) = \mathbf{1}[i,j \in S]/\Pr(i,j \in S)$ and the entries of $b_{\bar{\Omega}}$ are the product of the variance bound and outcomes, $b_{\bar{\Omega}}(i,j) = b_{ij}\theta_i\theta_j$. Note that the vector $z_{\bar{\Omega}}$ is random, while $b_{\bar{\Omega}}$ is fixed. Using the operator norm bound on the above, we have that the mean squared error may be bounded as

$$\mathbb{E}\Big[(\mathrm{VB}(\boldsymbol{\theta}) - \widehat{\mathrm{VB}}(\boldsymbol{\theta}))^2\Big] = b_{\bar{\Omega}}^{\intercal}\,\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})b_{\bar{\Omega}} \leq \Big\|\mathrm{Cov}(\boldsymbol{R}_{\bar{\Omega}})\Big\|_* \cdot \|b_{\bar{\Omega}}\|^2 \ .$$

Finally, we bound the squared ℓ_2 norm of the vector $b_{\bar{\Omega}}$. Using Hölder's inequality,

we have that for any integers p and q with $1/p + 1/q = 1$,

$$\|b_{\bar{\Omega}}\|^2 = \sum_{(i,j)\in\bar{\Omega}} b_{ij}^2(\theta_i\theta_j)^2 = \sum_{i=1}^{2n}\sum_{j=1}^{2n} b_{ij}^2(\theta_i\theta_j)^2 \leq \Big(\sum_{i=1}^{2n}\sum_{j=1}^{2n} b_{ij}^{2p}\Big)^{1/p}\Big(\sum_{i=1}^{2n}\sum_{j=1}^{2n}(\theta_i\theta_j)^{2q}\Big)^{1/q} .$$

We remark that the first term on the right hand side is the entry-wise $L_{2p,2}$ matrix norm, i.e. $\|\boldsymbol{B}\|_{2p,2}^2$. Additionally, by distributing terms, the second term on the right hand side may be simplified to $(\sum_{i=1}^{2n}\theta_i^{2q})^{2/q}$. □

We remark that Proposition 3.13 is obtained by setting $p = 1$ and $q = \infty$. By taking $q = 1$ and $p = \infty$, we get that the last term on the right hand side is the square of the second moment of the potential outcomes, which is bounded by the fourth moment of the potential outcomes via Jensen's inequality. Taking $q = 1$ and $p = \infty$, the second term on the right hand side becomes $\|\boldsymbol{B}\|_{\infty,2}^2$, which is the sum of the squares of the largest coefficient in each row of the matrix $\boldsymbol{B}$.

Using Proposition B.1, we may now establish more general conditions under which consistent estimation of the variance bound is possible.

Corollary B.2. *Suppose that the variance bound $\boldsymbol{B}$ is design compatible, $\|\mathrm{Cov}(z_{\bar{\Omega}})\|_*$ is bounded by a constant, and there exists integers p and q with $1/p+1/q = 1$ such that $\Big(\sum_{i=1}^{2n}\theta_i^{2q}\Big)^{2/q}$ is bounded by a constant. If $\|\boldsymbol{B}\|_{2p,2}^2 \to 0$ in the asymptotic sequence, then the Horvitz–Thompson estimator is a consistent estimator of the variance bound: $\mathbb{E}[(\,VB(\boldsymbol{\theta}) - \widehat{VB}(\boldsymbol{\theta}))^2] \to 0$.*

Motivated by Corollary B.2, some experimenters may wish to modify the regularized objective presented in Section 3.2.2 by replacing the square of the Frobenius norm with the square of the entry-wise $L_{2p,2}$ norm. Because the the square of the entry-wise $L_{2p,2}$ norm is convex, the resulting program will be convex and thus efficiently solvable with standard techniques. However, we do not know whether the the entry-wise $L_{2p,2}$ norm is monotone increasing with respect to the Loewner order and thus, the returned variance bound may or may not be admissible. We conjecture that the entry-wise $L_{2p,2}$ norm is monotone increasing with respect to the Loewner order and thus an admissible variance bound is returned.

B.3 Analysis of the Aronow–Samii bound

In this section, we demonstrate that the Aronow–Samii bound is admissible in certain experimental settings. In particular, the following proposition shows that in experimental settings where all second order pairwise assignments have nonzero probability of being observed, then a generalization of the Aronow–Samii bound is admissible, as it is returned by OPT-VB

Proposition B.3. *Consider the no-interference setting where* $\Pr(z_i = v_i, z_j = v_j) > 0$ *for all units* $i \neq j$ *and assignment values* $v_i, v_j \in \{0, 1\}$. *Suppose that* OPT-VB *is run with the linear objective* $g(\boldsymbol{S}) = \langle \boldsymbol{W}, \boldsymbol{S} \rangle$ *where* $\boldsymbol{W}$ *is a diagonal matrix with positive entries. Then, the resulting variance bound* $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{S}$ *is the following generalized Aronow–Samii bound:* $\boldsymbol{S} = \boldsymbol{S} = \frac{1}{2} \sum_{(k,\ell) \in \Omega} \boldsymbol{M}_{k\ell}$ *where* $\boldsymbol{M}_{k\ell}$ *be a* $2n \times 2n$ *matrix with zeros entries except in the* (k, ℓ)*th block, which is given by*

$$\begin{array}{c} \\ k \\ \ell \end{array} \begin{array}{c} \begin{array}{cc} k & \ell \end{array} \\ \begin{pmatrix} |a_{k\ell}| \left(\frac{w_{\ell\ell}}{w_{kk}}\right)^{1/2} & -a_{k\ell} \\ -a_{k\ell} & |a_{k\ell}| \left(\frac{w_{kk}}{w_{\ell\ell}}\right)^{1/2} \end{pmatrix} \end{array}.$$

Proof. To show that $\boldsymbol{S}^*$ is optimal, we will prove it is feasible and calculate its objective value. Then, we will show that any other feasible solution $\boldsymbol{S}$ has objective value which is at least that of $\boldsymbol{S}^*$.

Before continuing, we remark on the structure of Ω in this experimental setting. Because $\Pr(z_i = v_i, z_j = v_j) > 0$ for all units $i \neq j$ and assignment values $v_i, v_j \in \{0, 1\}$ and we assume no interference, then the pair of unobserved entries are given as

$$\Omega = \{(i, i + n) : i = 1, \ldots n\} \ .$$

Let us first show that $\boldsymbol{S}^*$ is a feasible solution. By construction, $\boldsymbol{S}^*$ is positive semidefinite and the off-diagonal entries satisfy $\boldsymbol{S}^*_{k\ell} = -a_{k\ell}$ for $(k, \ell) \in \Omega$. Now we must show that $\boldsymbol{S}^*$ is positive semidefinite. Because the sum of positive semidefinite matrices is positive semidefinite, it suffices to show that all of the individual matrices $\boldsymbol{M}_{k\ell}$ for $(k, \ell) \in \Omega$ are positive semidefinite. Fix a pair of unobserved entries $(k, \ell) \in \Omega$. Recall that because $\boldsymbol{W}$ is positive definite, all diagonal entries $w_{11}, w_{22} \ldots w_{2n,2n}$ are positive. The quadratic form in the matrix $\boldsymbol{M}_{k\ell}$ is non-negative by Young's inequality: given a $2n$-length vector $\boldsymbol{v}$,

$$\begin{aligned} \boldsymbol{v}^\intercal \boldsymbol{M}_{k\ell} \boldsymbol{v} &= \boldsymbol{v}_k^2 |a_{k\ell}| \Big(\frac{w_{\ell\ell}}{w_{kk}}\Big)^{1/2} + \boldsymbol{v}_\ell^2 |a_{k\ell}| \Big(\frac{w_{kk}}{w_{\ell\ell}}\Big)^{1/2} - 2a_{k\ell} \boldsymbol{v}_k \boldsymbol{v}_\ell \\ &\geq 2\Big(\boldsymbol{v}_k^2 |a_{k\ell}| \Big(\frac{w_{\ell\ell}}{w_{kk}}\Big)^{1/2} \boldsymbol{v}_\ell^2 |a_{k\ell}| \Big(\frac{w_{kk}}{w_{\ell\ell}}\Big)^{1/2}\Big)^{1/2} - 2a_{k\ell} \boldsymbol{v}_k \boldsymbol{v}_\ell \\ &= 2|a_{k\ell} \boldsymbol{v}_k \boldsymbol{v}_\ell| - 2a_{k\ell} \boldsymbol{v}_k \boldsymbol{v}_\ell \\ &\geq 0 \end{aligned}$$

Thus, we have shown that $\boldsymbol{S}^*$ is a feasible solution. Using the fact that $\boldsymbol{W}$ is diagonal,

we may calculate the objective value at $\boldsymbol{S}^*$ to be

$$\begin{aligned}
\langle \boldsymbol{W}, \boldsymbol{S}^* \rangle &= \sum_{k=1}^{2n} \sum_{\ell=1}^{2n} s^*_{k\ell} w_{k\ell} \\
&= \sum_{k=1}^{2n} s^*_{kk} w_{kk} \\
&= \sum_{(k,\ell)\in\Omega} \left(s^*_{kk} w_{kk} + s^*_{\ell\ell} w_{\ell\ell} \right) \\
&= \sum_{(k,\ell)\in\Omega} \left(|a_{k\ell}| \Big(\frac{w_{\ell\ell}}{w_{kk}} \Big)^{1/2} w_{kk} + |a_{k\ell}| \Big(\frac{w_{kk}}{w_{\ell\ell}} \Big)^{1/2} w_{\ell\ell} \right) \\
&= 2 \sum_{(k,\ell)\in\Omega} |a_{k\ell}| (w_{kk} w_{\ell\ell})^{1/2} \ .
\end{aligned}$$

Let $\boldsymbol{S}$ be any feasible solution. Let $(k, \ell) \in \Omega$ be given. Because $\boldsymbol{S}$ is positive semidefinite, we have that the diagonal entries are non-negative, $s_{kk} \geq 0$. In addition, the off-diagonal entries satisfy the inequality

$$s_{kk} s_{\ell\ell} \geq s^2_{k\ell} = a^2_{k\ell} \ .$$

Together, these demonstrate that its objective value may be lower bounded as

$$\begin{aligned}
\langle \boldsymbol{W}, \boldsymbol{S} \rangle &= \sum_{k=1}^{2n} \sum_{\ell=1}^{2n} s_{k\ell} w_{k\ell} \\
&= \sum_{k=1}^{2n} s_{kk} w_{kk} && (\boldsymbol{W} \text{ is diagonal}) \\
&= \sum_{(k,\ell)\in\Omega} s_{kk} w_{kk} + s_{\ell\ell} w_{\ell\ell} && (\text{assumptions on } \Omega) \\
&\geq \sum_{(k,\ell)\in\Omega} 2 (s_{kk} s_{\ell\ell} w_{kk} w_{\ell\ell})^{1/2} && (\text{for } a, b \geq 0,\ a + b \geq 2(ab)^{1/2}) \\
&\geq \sum_{(k,\ell)\in\Omega} 2 (a^2_{k\ell} w_{kk} w_{\ell\ell})^{1/2} && (s_{kk} s_{\ell\ell} \geq a^2_{k\ell}) \\
&= 2 \sum_{(k,\ell)\in\Omega} |a_{k\ell}| (w_{kk} w_{\ell\ell})^{1/2} \\
&= \langle \boldsymbol{W}, \boldsymbol{S}^* \rangle
\end{aligned}$$

so that $\langle \boldsymbol{W}, \boldsymbol{S} \rangle \geq \langle \boldsymbol{W}, \boldsymbol{S}^* \rangle$ for all feasible $\boldsymbol{S}$. Thus, $\boldsymbol{S}^*$ is the optimal solution to the program underlying OPT-VB, as desired. □

Proposition B.3 provides us with the targeting objective matrix used to derive generalized Aronow–Samii bounds. As proposed in Section 3.3.5, we can re-interpret

these variance bounds by interpreting the objective matrix to encode the raw second moment of the potential outcomes under a generative model, i.e. $\boldsymbol{W} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal]$. For example, a generative model which assumes that outcomes are uncorrelated and mean zero would result in $\boldsymbol{W} = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta}\boldsymbol{\theta}^\intercal]$ being diagonal. In particular, the diagonal entries correspond to the variances of each individual outcome. In this way, we may re-interpret the Aronow-Samii bound (in the no-interference settings where it is admissible) as minimizing the expectation of the variance bound under in a generative model where outcomes are uncorrelated and mean zero.

Appendix C

Appendix for Bipartite Experiments Under a Linear Exposure-Response Assumption

C.1 Analysis of the ERL Estimator

In this section, we present proofs of consistency and asymptotic normality of ERL estimator appearing in Section 4.3 of Chapter 4. Before continuing, we introduce some notation used in the proofs. We begin by defining for each outcome unit $i \in V_o$, an estimate of the individual treatment effect τ_i, which is

$$\hat{\tau}_i \triangleq 2y_i(\boldsymbol{z})\Big(\frac{x_i(\boldsymbol{z}) - \mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\Big) \ .$$

Observe that the ERL estimator is the average of these estimates of the individual treatment effects, i.e. $\hat{\tau} = (1/n)\sum_{i=1}^{n} \hat{\tau}_i$. Throughout the proofs, we will often reason about the behavior of the ERL estimator through the properties of the individual treatment effect estimates.

Next, we introduce the concept of *dependency neighborhoods* (Ross, 2011). Let $a_1, a_2, \ldots a_n$ be random variables indexed by the integers $[n]$ and collect these random variables into the set $\mathcal{A} = \{a_i : i \in [n]\}$. For each variable a_i, we define the *dependency neighborhood* as

$$\mathcal{I}(i) \subset \mathcal{A} \text{ such that } a_i \text{ is jointly independent of the variables } \mathcal{A} \setminus \mathcal{I}(i) \ .$$

In other words, a random variable a_i is jointly independent of all variables not contained in its dependency neighborhood, but is dependent on variables contained in its dependency neighborhood. We take the convention that $i \in \mathcal{I}(i)$ and so that each dependency neighborhood as cardinality at least 1. A measure of dependence between the random variables is the *maximum dependency degree*, which is $D = \max_{i\in[n]}|\mathcal{I}(i)|$. Note that independent random variables satisfy $D = 1$ and that completely dependent

random variables have $D = n$.

For the remainder of the proof, we focus our discussion of dependency neighborhoods and degrees to the collection of errors of the individual treatment effects,

$$a_1 = \tau_1 - \hat{\tau}_1, \quad a_2 = \tau_2 - \hat{\tau}_2, \quad \ldots \quad a_n = \tau_n - \hat{\tau}_n \ .$$

We begin by showing that in this case, the maximum dependency degree may be bounded in terms of the degrees of the bipartite graph and the dependence in the treatment assignments.

Lemma C.1. *The dependency degree of the individual treatment effect errors is bounded by* $D \leq k d_d d_o$.

Proof. The first part of this proof is to establish a necessary condition for an individual treatment effect error a_j to be in the dependency neighborhood of a_i, i.e. $a_j \in \mathcal{I}(i)$. We begin by re-writing the exposures under a cluster design. Recall that the exposures are defined as $x_i = \sum_{j=1}^{m} w_{i,j} z_j$. For each cluster $C \in \mathcal{C}$, define $w_{i,C} = \sum_{j \in C} w_{i,j}$ and define z_C to be the ± 1 cluster treatment assignment variable which is 1 if diversion units in C are treated and -1 otherwise. If $w_{i,C} \neq 0$, then we say that cluster C is *incident* to outcome unit i. Define $S(i) = \{z_C : w_{i,C} \neq 0\}$ to be the cluster treatment assignments which influence the exposure x_i. Under the cluster design, the exposure for outcome unit i may be written as

$$x_i = \sum_{C \in \mathcal{C}} w_{i,C} z_C = \sum_{C \in S(i)} w_{i,C} z_C \ .$$

By the linear-response assumption, the individual treatment effect error a_i is a function of the exposure x_i. Moreover, a_i is a function of the cluster treatment assignment variables in $S(i)$. Let us denote this relationship by writing $a_i = g_i(S(i))$, where g_i is a function of the cluster treatment variables $z_C \in S(i)$. Let $B \subset V_o$ be a collection of outcome units. We remark that joint independence of cluster treatment assignments implies joint independence of individual treatment effect errors:

$$S(i) \perp\!\!\!\perp \{S(j) : j \in B\} \Rightarrow a_i \perp\!\!\!\perp \{a_j : j \in B\} \ .$$

Under an independent cluster design, the cluster treatment assignments $S(i)$ are jointly independent of the cluster treatment assignments $\{S(j) : j \in B\}$ when the corresponding sets of clusters are disjoint, i.e. $S(i) \cap (\cup_{j \in B} S(j)) = \emptyset$. Thus, the individual treatment effect estimate a_i is jointly independent of the collection of individual treatment effect estimates $\{a_j : j \in B\}$ when outcome unit i is not incident to any cluster that is incident to an outcome unit in B. In other words, $a_j \in \mathcal{I}(i)$ only if outcome units i and j are incident to a common cluster.

Fix an outcome unit $i \in V_o$. The remainder of the proof is a simple counting argument which uses this necessary condition to establish that $|\mathcal{I}(i)| \leq k d_d d_o$. In particular, we will count the number of outcome units that are incident to one of the

clusters that are incident to i. Because the degree of outcome unit i is at most d_o, it is incident to at most d_o clusters. Each of these clusters has at most k diversion units, by Assumption 4.4. Because the degree of all diversion units j is at most d_d, the number of outcome units which are incident to at least one of these clusters is at most kd_dd_o. Thus, we have established that

$$D = \max_{i \in V_o} |\mathcal{I}(i)| \leq kd_dd_o \ . \qquad \square$$

The following lemma derives a lower bound the exposure variances in terms of the treatment assignment probability and the maximum degree of the outcome units.

Lemma C.2. *If each pair of treatment assignments is non-negatively correlated, then each exposure variance is lower bounded as* $\mathrm{Var}(x_i) \geq \frac{4p(1-p)}{d_o}$.

Proof. We begin by expanding the variance of the exposure x_i by

$$\begin{aligned}
\mathrm{Var}(x_i) &= \mathrm{Var}\Big(\sum_{j=1}^{m} w_{i,j} z_j\Big) && \text{(definition of exposure)} \\
&= \sum_{i=1}^{m} \Big[\mathrm{Var}(w_{i,j} z_j) + \sum_{\ell \neq j} \mathrm{Cov}(w_{i,j} z_j, w_{i,\ell} z_\ell)\Big] && \text{(properties of Var)} \\
&= \sum_{i=1}^{m} \Big[w_{i,j}^2 \,\mathrm{Var}(z_j) + \sum_{\ell \neq j} w_{i,j} w_{i,\ell} \,\mathrm{Cov}(z_j, z_\ell)\Big] && \text{(properties of Var and Cov)} \\
&\geq \sum_{i=1}^{m} w_{i,j}^2 \,\mathrm{Var}(z_j) && \text{(non-negativity)} \\
&= 4p(1-p) \sum_{i=1}^{m} w_{i,j}^2 \ ,
\end{aligned}$$

where the inequality follows because the weights $w_{i,j}$ are non-negative and the assignments are non-negatively correlated and the last equality follows because z_i are ± 1 random variables with $\Pr(z_i = 1) = p$.

We complete the proof by lower bounding the sum of the squares of the weights. Recall that the sum of the weights is 1 and there are at most d_o non-negative terms in the sum. Using this together with the inequality that relates ℓ_2 to ℓ_1 norms in d-dimensions, $\|\cdot\|_2^2 \geq \frac{1}{d}\|\cdot\|_1^2$, we have that

$$\sum_{i=1}^{m} w_{i,j}^2 \geq \frac{1}{d_o} \sum_{i=1}^{m} w_{i,j} = \frac{1}{d_o} \ . \qquad \square$$

The following lemma is a bound on the moments of the errors of the individual treatment effects.

Lemma C.3. *The pth moment of the error of the individual treatment effect estimates is bounded by*

$$\mathbb{E}[|\tau_i - \hat{\tau}_i|^p] \leq \left[2M\Big(1 + \frac{d_o}{2p(1-p)}\Big)\right]^p .$$

Proof. We begin by remarking that $|y_i(\boldsymbol{z})| \leq M$ implies that each of the individual slopes are also bounded in absolute value as $|\beta_i| \leq M$. Recall that by the linear response assumption, $y_i(\boldsymbol{z}) = \beta_i x_i + \alpha_i$ and by the linear exposure assumption (along with the normalization of the edge weights), setting $\boldsymbol{z} = \pm\mathbf{1}$ results in an exposure of $\xi = \pm 1$. Thus, when considering $\boldsymbol{z} = \pm\mathbf{1}$, the bound $|y_i(\boldsymbol{z})| \leq M$ implies that $|\pm\beta_i + \alpha_i| \leq M$, which is enough to establish that $|\beta_i| \leq M$.

We now proceed by proving a bound on $|\tau_i - \hat{\tau}_i|$, which holds for any realization of the random variables:

$$\begin{aligned}
|\tau_i - \hat{\tau}_i| &= \Big|2\beta_i - 2y_i(\boldsymbol{z})\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big| \\
&\leq 2|\beta_i| + 2\frac{|y_i(\boldsymbol{z})| \cdot |x_i - \mathbb{E}[x_i]|}{\mathrm{Var}(x_i)} && \text{(triangle inequality)} \\
&\leq 2M + 2\frac{M \cdot 2}{\mathrm{Var}(x_i)} && \text{(definition of } M \text{ and above)} \\
&\leq 2M\Big(1 + \frac{2}{\mathrm{Var}(x_i)}\Big) && \text{(collecting terms)} \\
&\leq 2M\Big(1 + \frac{d_o}{2p(1-p)}\Big) && \text{(Lemma C.2)}
\end{aligned}$$

The moment bound follows by applying the bound above. □

C.1.1 Expectation of the ERL estimator (Theorems 4.2 and 4.11)

In this section, we derive the expectation of the ERL estimator, both with and without the linear exposure-response assumption. First, we derive the expectation under the linear exposure-response assumption.

Theorem 4.2. *Suppose the design is such that each exposure has a positive variance. Under the linear response assumption, the* ERL *estimator is unbiased for the ATTE:* $\mathbb{E}[\hat{\tau}] = \tau$.

Proof. By linearity, the expectation of the estimator is

$$\mathbb{E}[\hat{\tau}] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}\left[y_i(\boldsymbol{z})\left(\frac{x_i(\boldsymbol{z}) - \mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] .$$

By Proposition 4.1, the ATTE is twice the average of the slope terms β_i. Thus, to complete the proof we show that each expectation terms inside the sum is equal to

the corresponding slope β_i. Using the linear response assumption,

$$\begin{aligned}
&\mathbb{E}\left[y_i(\boldsymbol{z})\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] \\
&= \mathbb{E}\left[(\beta_i x_i(\boldsymbol{z})+\alpha_i)\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] \\
&= \beta_i\,\mathbb{E}\left[x_i(\boldsymbol{z})\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] + \alpha_i\,\mathbb{E}\left[\left(\frac{x_i(\boldsymbol{z})-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right)\right] \\
&= \beta_i\left(\frac{\mathbb{E}[x_i(\boldsymbol{z})^2]-\mathbb{E}[x_i(\boldsymbol{z})]^2}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right) + \alpha_i\left(\frac{\mathbb{E}[x_i(\boldsymbol{z})]-\mathbb{E}[x_i(\boldsymbol{z})]}{\mathrm{Var}(x_i(\boldsymbol{z}))}\right) \\
&= \beta_i
\end{aligned}$$

□

Next, we derive the expectation of the ERL estimator under a general (non-linear) response assumption.

Theorem 4.11. *Assume that the potential functions are an arbitrary function of the exposures: $y_i(\boldsymbol{z}) = y_i(x_i)$. Then, the expectation of the* ERL *estimator is*

$$\mathbb{E}[\widehat{\tau}] = \frac{2}{n}\sum_{i=1}^{n}\hat{\beta}_i \ ,$$

where $\hat{\beta}_i$ is the coefficient of the exposure x_i in an OLS regression of y_i on x_i: $\hat{\beta}_i = \left(\frac{\mathrm{Cov}(x_i, y_i(x_i))}{\mathrm{Var}(x_i)}\right)$.

Proof. We begin by deriving the expectation of an individual term in the ERL estimator. To this end, observe that

$$\mathbb{E}\Big[y_i\Big(\frac{x_i-\mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big] = \frac{\mathbb{E}[y_i x_i]-\mathbb{E}[y_i]\,\mathbb{E}[x_i]}{\mathrm{Var}(x_i)} = \frac{\mathrm{Cov}(x_i, y_i)}{\mathrm{Var}(x_i)} \ .$$

The proof is completed by linearity of expectation. □

We remark that Theorem 4.2 follows from Theorem 4.11 by observing that under a linear response assumption that $y_i = \beta_i x_i + \alpha_i$, we have that $\mathrm{Cov}(x_i, y_i) = \mathrm{Cov}(x_i, \beta_i x_i + \alpha_i) = \beta_i \,\mathrm{Var}(x_i)$.

C.1.2 Consistency of ERL estimator (Theorem 4.5)

We are now ready to establish the consistency of the ERL estimator. Before doing so, we restate the theorem here.

Theorem 4.5. *Under Assumptions 4.3 and 4.4, and supposing that $d_d d_o^3 = o(n)$ in the asymptotic sequence, the* ERL *estimator converges in mean square to the ATTE: $\lim_{N\to\infty}\mathbb{E}[(\hat{\tau}-\tau)^2] = 0$.*

Proof. We begin by proving a finite sample bound on the mean squared error of the ERL estimator, and then we finish the proof by taking the limit in the asymptotic sequence. Note that the mean squared error may be broken down into the errors of the individual treatment effect estimates via

$$\mathbb{E}[(\tau-\hat{\tau})^2] = \mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}(\tau_i-\hat{\tau}_i)\Big)^2\Big] = \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathbb{E}[(\tau_i-\hat{\tau}_i)^2]+\sum_{j\neq i}\mathbb{E}[(\tau_i-\hat{\tau}_i)(\tau_j-\hat{\tau}_j)]\right) .$$

Note that the term in the inner sum is the covariance of the errors in the individual treatment effect estimators. By definition of the dependency neighborhoods, only terms $j \in \mathcal{I}(i)$ are dependent and so only these terms will have non-zero covariance. Using this and the second moment bound in Lemma C.3, we have that

$$\begin{aligned}
\mathbb{E}[(\tau-\hat{\tau})^2] &= \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathbb{E}[(\tau_i-\hat{\tau}_i)^2]+\sum_{\substack{j\in\mathcal{I}(i)\\ j\neq i}}\mathbb{E}[(\tau_i-\hat{\tau}_i)(\tau_j-\hat{\tau}_j)]\right)\\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathbb{E}[(\tau_i-\hat{\tau}_i)^2]+\sum_{\substack{j\in\mathcal{I}(i)\\ j\neq i}}\sqrt{\mathbb{E}[(\tau_i-\hat{\tau}_i)^2]\,\mathbb{E}[(\tau_j-\hat{\tau}_j)^2]}\right)\\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}|\mathcal{I}(i)|\cdot\Big[2M\Big(1+\frac{d_o}{2p(1-p)}\Big)\Big]^2\\
&\leq \frac{D}{n}\Big[2M\Big(1+\frac{d_o}{2p(1-p)}\Big)\Big]^2 ,
\end{aligned}$$

where the first equality holds by the definition of dependency neighborhoods, the first inequality is Cauchy-Schwarz, the second inequality follows from Lemma C.3, and the final inequality follows from the bound on the maximum dependency degree. By using the bound $D \leq kd_dd_o$ given in Lemma C.1, we have the finite-sample bound on the mean squared error:

$$\mathbb{E}[(\tau-\hat{\tau})^2] \leq \frac{kd_dd_o}{n}\Big[2M\Big(1+\frac{d_o}{2p(1-p)}\Big)\Big]^2$$

We now interpret this finite-sample bound in the context of the asymptotic sequence. By Assumptions 4.3 and 4.4, we have that M is a constant, p is bounded away from 0 and 1 by a constant, and k is a constant. It follows that the mean squared error is asymptotically bounded by the rate $\mathbb{E}[(\tau-\hat{\tau})^2] = \mathcal{O}(d_dd_o^3/n)$. By assumption, the asymptotic sequence satisfies $d_dd_o^3 = o(n)$, and thus the mean squared error converges to zero under these conditions. □

C.1.3 Asymptotic normality (Theorem 4.7)

We establish asymptotic normality of the ERL estimator by using Stein's method. In particular, we use the following result from Ross (2011):

Lemma C.4 (Lemma 3.6 of Ross (2011))**.** *Let $a_1, a_2, \dots a_n$ be random variables such that $\mathbb{E}[a_i^4] < \infty$, $\mathbb{E}[a_i] = 0$, $\sigma^2 = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^n a_i)$, and define $X = (\frac{1}{n}\sum_{i=1}^n a_i)/\sigma$. Then for a standard normal $Z \sim \mathcal{N}(0,1)$, we have*

$$d_W(X,Z) \leq \frac{D^2}{\sigma^3 n^3}\sum_{i=1}^n \mathbb{E}[|a_i|^3] + \sqrt{\frac{28}{\pi}}\cdot\frac{D^{3/2}}{n^2\sigma^2}\sqrt{\sum_{i=1}^n \mathbb{E}[a_i^4]}\ ,$$

where D is the maximum dependency degree of the random variables and $d_W(\cdot,\cdot)$ is the Wasserstein distance.

We will use Lemma C.4 to prove asymptotic normality of the ERL estimator. Before continuing, let us restate the theorem.

Theorem 4.7. *Under Assumptions 4.3, 4.4, and 4.6, and supposing that $d_d^{1.6} d_o^4 = o(n)$, the* ERL *estimator is asymptotically normal:*

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

Proof. Our strategy may be described in two main steps: first, we use Lemma C.4 to derive a finite-sample bound on the Wasserstein distance between the distribution of $(\tau - \hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$ and a standard normal. Next, we us this bound to argue that this Wasserstein distance approaches 0 in the limit of the asymptotic sequence under the above conditions.

We seek to apply Lemma C.4 where the random variables are the errors of the individual treatment effect estimates; that is,

$$a_1 = \tau_1 - \hat{\tau}_1, a_2 = \tau_2 - \hat{\tau}_2, \dots a_n = \tau_n - \hat{\tau}_n\ .$$

Note that $\frac{1}{n}\sum_{i=1}^n a_i = \frac{1}{n}\sum_{i=1}^n \tau_i - \hat{\tau}_i = \tau - \hat{\tau}$ and $\mathrm{Var}(\frac{1}{n}\sum_{i=1}^n a_i) = \mathrm{Var}(\hat{\tau})$ so that the random variable X in Lemma C.4 is equal to $(\tau - \hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$, which is indeed the random variable we wish to characterize. Let's show that the conditions of Lemma C.4 are satisfied: first, recall that $\hat{\tau}_i$ are unbiased estimates of τ_i so that a_i has mean zero. Second, because the potential outcomes are bounded by a constant M, the support of a_i is bounded so the fourth moments are finite. Thus, we may apply Lemma C.4 in this setting.

We will use the Lemma C.3 to bound the sum of the third and fourth moments.

In particular, Lemma C.3 implies that

$$\sum_{i=1}^{n} \mathbb{E}[|a_i|^3] \leq n \cdot \left[2M\Big(1+\frac{d_o}{2p(1-p)}\Big)\right]^3 \text{ and } \sum_{i=1}^{n} \mathbb{E}[|a_i|^4] \leq n \cdot \left[2M\Big(1+\frac{d_o}{2p(1-p)}\Big)\right]^4 .$$

Using this moment bound together with the bound on the maximum dependence degree D (Lemma C.1) on the result of Lemma C.4, we obtain that the Wasserstein distance $d_W\Big(\frac{\tau-\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}, Z\Big)$ is at most

$$\frac{(kd_d d_o)^2}{\sigma^3 n^2} \cdot \left[2M\Big(1+\frac{d_o}{2p(1-p)}\Big)\right]^3 + \sqrt{\frac{28}{\pi}} \cdot \frac{(kd_d d_o)^{3/2}}{\sigma^2 n^{3/2}} \cdot \left[2M\Big(1+\frac{d_o}{2p(1-p)}\Big)\right]^2 .$$

We now interpret this finite-sample bound in the context of the asymptotic sequence. By Assumptions 4.3 and 4.4, we have that M is a constant, p is bounded away from 0 and 1 by a constant, and k is a constant. It follows that the Wasserstein distance between $(\tau-\hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$ and a standard normal is asymptotically bounded as

$$d_W\Big(\frac{\tau-\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}, Z\Big) = \mathcal{O}\Big(\frac{d_d^2 d_o^5}{\sigma^3 n^2} + \frac{d_d^{1.5} d_o^{3.5}}{\sigma^2 n^{3/2}}\Big)$$

By Assumption 4.6, we have that $\mathrm{Var}(\hat{\tau}) = \Omega(n^{-1/2})$, which means that this bound becomes

$$d_W\Big(\frac{\tau-\hat{\tau}}{\sqrt{\mathrm{Var}(\hat{\tau})}}, Z\Big) = \mathcal{O}\Big(\frac{d_d^2 d_o^5}{n^{5/4}} + \frac{d_d^{1.5} d_o^{3.5}}{n}\Big) .$$

By assumption, the asymptotic sequence satisfies $d_d^{1.6} d_o^4 = o(n)$. Raising both sides to the 5/4 yields that $d_d^2 d_o^5 = o(n^{5/4})$. Additionally, d_d and d_o are positive integers and so they satisfy $d_d^{1.5} d_o^{3.5} \leq d_d^{1.6} d_o^4$, thus $d_d^{1.5} d_o^{3.5} = o(n)$. Thus, the Wasserstein distance between $(\tau-\hat{\tau})/\sqrt{\mathrm{Var}(\hat{\tau})}$ and a standard normal approaches 0 in this asymptotic sequence. □

C.2 Variance Estimation

In this section, we prove Lemmas 4.8 and 4.9, which are then used to prove Theorem 4.10, which establishes unbiasedness of the proposed variance estimator.

First, we show Lemma 4.8, which demonstrates how to estimate the variance of an individual treatment effect estimate. To do so, we need to decompose the random variable Q_i. For each outcome unit $i \in V_o$, define the random variable T_i to be the second term in the function Q_i. That is,

$$T_i = \frac{\mathrm{Var}(x_i)(x_i^2 - \mathbb{E}[x_i^2]) - \mathrm{Cov}(x_i, x_i^2)(x_i - \mathbb{E}[x_i])}{\mathrm{Var}(x_i)\,\mathrm{Var}(x_i^2) - \mathrm{Cov}(x_i, x_i^2)^2}$$

so that $Q_i \triangleq \frac{(x_i - \mathbb{E}[x_i])^2}{\mathrm{Var}(x_i)^2} - T_i$. The following Lemma states the key properties of the random variable T_i, and consequently Q_i, which allow for the proof of Lemma 4.8.

Lemma C.5. *If the exposure x_i takes at least three values with non-zero probability, then the function T_i satisfies the following three properties:*

- $\mathbb{E}[T_i] = 0$
- $\mathbb{E}[x_i T_i] = 0$
- $\mathbb{E}[x_i^2 T_i] = 1$

Proof. For notational simplicity, we drop the subscript as write x_i as x and T_i as T.

First, we will show that the denominator $\mathrm{Var}(x)\,\mathrm{Var}(x^2) - \mathrm{Cov}(x, x^2)^2$ is positive when the exposure x takes at least three values with non-zero probability. By Cauchy-Schwarz, we have that $\mathrm{Cov}(x, x^2)^2 \leq \mathrm{Var}(x)\,\mathrm{Var}(x^2)$, which establishes that denominator is non-negative. The Cauchy-Schwarz inequality is strict precisely when x and x^2 are not perfectly correlated, i.e. there does not exist a and b such that $x^2 = ax + b$. Note that this cannot happen when x takes three distinct values. Thus, the Cauchy-Schwarz is strict in this case so that the denominator is positive. For notational convenience, we write the denominator as $\Delta = \mathrm{Var}(x)\,\mathrm{Var}(x^2) - \mathrm{Cov}(x, x^2)^2$ throughout the remainder of the proof.

We will now show the three properties. The first property follows by linearity of expectation, as

$$\begin{aligned}
\Delta \cdot \mathbb{E}[T] &= \mathbb{E}[\mathrm{Var}(x_i)(x^2 - \mathbb{E}[x^2]) - \mathrm{Cov}(x, x^2)(x - \mathbb{E}[x])] \\
&= \mathrm{Var}(x)(\mathbb{E}[x^2] - \mathbb{E}[x^2]) - \mathrm{Cov}(x, x^2)(\mathbb{E}[x] - \mathbb{E}[x]) \\
&= 0 \ ,
\end{aligned}$$

and the result follows by dividing through by Δ on both sides. Next, we show the second property. Again, by linearity of expectation, we have that

$$\begin{aligned}
\Delta \cdot \mathbb{E}[xT] &= \mathbb{E}[\mathrm{Var}(x_i)(x^3 - x\,\mathbb{E}[x^2]) - \mathrm{Cov}(x, x^2)(x^2 - x\,\mathbb{E}[x])] \\
&= \mathrm{Var}(x)(\mathbb{E}[x^3] - \mathbb{E}[x]\,\mathbb{E}[x^2]) - \mathrm{Cov}(x, x^2)(\mathbb{E}[x^2] - \mathbb{E}[x]^2) \\
&= \mathrm{Var}(x)\,\mathrm{Cov}(x, x^2) - \mathrm{Cov}(x, x^2)\,\mathrm{Var}(x) \\
&= 0 \ ,
\end{aligned}$$

and the result follows by dividing through by Δ on both sides. Finally, we show the third property. By linearity of expectation,

$$\begin{aligned}
\Delta \cdot \mathbb{E}[x^2 T] &= \mathbb{E}[\mathrm{Var}(x_i)(x^4 - x^2\,\mathbb{E}[x^2]) - \mathrm{Cov}(x, x^2)(x^3 - x^2\,\mathbb{E}[x])] \\
&= \mathrm{Var}(x)(\mathbb{E}[x^4] - \mathbb{E}[x^2]^2) - \mathrm{Cov}(x, x^2)(\mathbb{E}[x^3] - \mathbb{E}[x^2]\,\mathbb{E}[x]) \\
&= \mathrm{Var}(x)\,\mathrm{Var}(x^2) - \mathrm{Cov}(x, x^2)^2 \\
&= \Delta \ ,
\end{aligned}$$

and the result follows by dividing both sides by Δ. $\square$

We are now ready to prove Lemma 4.8, which we restate here.

Lemma 4.8. *Fix an outcome unit $i \in V_o$. If the exposure x_i takes at least three values with non-zero probability, then the variance of unit i's individual treatment effect estimator is equal to*

$$\mathrm{Var}(\hat{\tau}_i) = 4 \cdot \mathbb{E}[y_i(\boldsymbol{z})^2 Q_i] \ .$$

Proof. We may use the properties of the random variable T_i proved in Lemma C.5 together with the linear response assumption to obtain

$$\begin{aligned}
\mathbb{E}\Big[y_i(\boldsymbol{z})^2 T_i\Big] &= \mathbb{E}\Big[(\beta_i x_i + \alpha_i)^2 T_i\Big] && \text{(linear response)} \\
&= \beta_i^2 \, \mathbb{E}\Big[x_i^2 T_i\Big] + 2\beta_i\alpha_i \, \mathbb{E}\Big[x_i T_i\Big] + \alpha_i^2 \, \mathbb{E}\Big[T_i\Big] && \text{(expanding terms)} \\
&= \beta_i^2. && \text{(Lemma C.5)}
\end{aligned}$$

Next, we write the variance of individual treatment effect estimator. Because the individual treatment effect estimator is unbiased,

$$\begin{aligned}
\mathrm{Var}(\hat{\tau}_i) &= \mathbb{E}[\hat{\tau}_i^2] - \mathbb{E}[\hat{\tau}_i]^2 \\
&= 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})^2 \Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)^2\Big] - 4 \cdot \beta_i^2 \\
&= 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})^2 \Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)^2\Big] - 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})^2 T_i\Big] \\
&= 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})^2 \Big[\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)^2 - T_i\Big]\Big] \\
&= 4 \cdot \mathbb{E}[y_i(\boldsymbol{z})^2 Q_i] \ .
\end{aligned}$$

$\square$

Next, we will prove Lemma 4.9 which demonstrates how to estimate the covariance between two individual treatment effects. To do so, we need to decompose the random variable $R_{i,j}$. For each pair of outcome units $i, j \in V_o$, define the random variable $S_{i,j}$ to be the second term in the random variable $R_{i,j}$. That is,

$$S_{i,j} = \frac{a_{i,j}(x_i x_j - \mathbb{E}[x_i x_j]) + b_{i,j}(x_i - \mathbb{E}[x_i]) + c_{i,j}(x_j - \mathbb{E}[x_j])}{\Psi_{i,j}} \ ,$$

where we recall that these coefficients are given as $a_{i,j}, b_{i,j}, c_{i,j}, \Psi_{i,j}$ are defined as

$$\begin{aligned}
a_{i,j} &= \mathrm{Var}(x_i)\,\mathrm{Var}(x_j) - \mathrm{Cov}(x_i, x_j)^2 \\
b_{i,j} &= \mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_j) - \mathrm{Var}(x_j)\,\mathrm{Cov}(x_i x_j, x_i) \\
c_{i,j} &= \mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_i) - \mathrm{Var}(x_i)\,\mathrm{Cov}(x_i x_j, x_j) \\
\Psi_{i,j} &= \mathrm{Var}(x_i x_j)(\mathrm{Var}(x_i)\,\mathrm{Var}(x_j) - \mathrm{Cov}(x_i, x_j)^2) - \mathrm{Var}(x_i)\,\mathrm{Cov}(x_i x_j, x_j)^2 \\
&\quad - \mathrm{Var}(x_j)\,\mathrm{Cov}(x_i x_j, x_i)^2 + 2\,\mathrm{Cov}(x_i, x_j)\,\mathrm{Cov}(x_i x_j, x_j)\,\mathrm{Cov}(x_i x_j, x_i) \ .
\end{aligned}$$

In this way, $R_{i,j} \triangleq (\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)})(\frac{x_j - \mathbb{E}[x_j]}{\mathrm{Var}(x_j)}) - S_{i,j}$. The following lemma states the key properties of the random variable $S_{i,j}$, and consequently $R_{i,j}$, which allow for the proof of Lemma 4.9.

Lemma C.6. *Fix a pair of outcome units $i, j \in V_o$. If $\Psi_{i,j} \neq 0$, then the random variable $S_{i,j}$ satisfies the following properties:*

- $\mathbb{E}[S_{i,j}] = 0$
- $\mathbb{E}[x_i S_{i,j}] = \mathbb{E}[x_j S_{i,j}] = 0$
- $\mathbb{E}[x_i x_j S_{i,j}] = 1$

Proof. For notational simplicity, we will drop the subscripts and write x_i as x, x_j as y and $S_{i,j}$ as S. In the same way, we will write $a_{i,j}, b_{i,j}, c_{i,j}$, and $\Psi_{i,j}$ as a, b, c, and Ψ.

Given that $\Psi > 0$, we seek to verify the three properties in the lemma. The first is easy to verify by the linearity of expectation, as

$$\begin{aligned}
\Psi \cdot \mathbb{E}[S] &= \mathbb{E}[a(xy - \mathbb{E}[xy]) + b(x - \mathbb{E}[x]) + c(y - \mathbb{E}[y])] \\
&= a(\mathbb{E}[xy] - \mathbb{E}[xy]) + b(\mathbb{E}[x] - \mathbb{E}[x]) + c(\mathbb{E}[y] - \mathbb{E}[y] \\
&= 0
\end{aligned}$$

so that dividing both sides by Ψ yields the desired result. We now verify the second property. Observe that

$$\begin{aligned}
\Psi \cdot \mathbb{E}[xS] &= a \cdot (\mathbb{E}[x^2 y] - \mathbb{E}[x]\,\mathbb{E}[xy]) + b \cdot (\mathbb{E}[x^2] - \mathbb{E}[x]^2) + c \cdot (\mathbb{E}[xy] - \mathbb{E}[x]\,\mathbb{E}[y]) \\
&= a \cdot \mathrm{Cov}(xy, x) + b \cdot \mathrm{Var}(x) + c \cdot \mathrm{Cov}(x, y) \\
&= \mathrm{Cov}(xy, x)(\mathrm{Var}(x)\,\mathrm{Var}(y) - \mathrm{Cov}(x, y)^2) \\
&\quad + \mathrm{Var}(x)(\mathrm{Cov}(x, y)\,\mathrm{Cov}(xy, y) - \mathrm{Var}(y)\,\mathrm{Cov}(xy, x)) \\
&\quad + \mathrm{Cov}(x, y)(\mathrm{Cov}(x, y)\,\mathrm{Cov}(xy, x) - \mathrm{Var}(x)\,\mathrm{Cov}(xy, y)) \\
&= 0 \ ,
\end{aligned}$$

and dividing both sides by Ψ yields the desired result. The proof that $\mathbb{E}[yS(x, y)] = 0$ follows in exactly the same fashion. Finally, we verify the third property. Observe

that

$$\begin{aligned}
\Psi \cdot \mathbb{E}[xyS] &= a \cdot (\mathbb{E}[(xy)^2] - \mathbb{E}[xy]^2) + b \cdot (\mathbb{E}[x^2 y] - \mathbb{E}[xy]\,\mathbb{E}[x]) \\
&\quad + c \cdot (\mathbb{E}[xy^2] - \mathbb{E}[xy]\,\mathbb{E}[y]) \\
&= a \cdot \mathrm{Var}(xy) + b \cdot \mathrm{Cov}(xy, x) + c \cdot \mathrm{Cov}(xy, y) \\
&= \mathrm{Var}(xy)(\mathrm{Var}(x)\,\mathrm{Var}(y) - \mathrm{Cov}(x, y)^2) \\
&\quad\quad + \mathrm{Cov}(xy, x)(\mathrm{Cov}(x, y)\,\mathrm{Cov}(xy, y) - \mathrm{Var}(y)\,\mathrm{Cov}(xy, x)) \\
&\quad\quad + \mathrm{Cov}(xy, y)(\mathrm{Cov}(x, y)\,\mathrm{Cov}(xy, x) - \mathrm{Var}(x)\,\mathrm{Cov}(xy, y)) \\
&= \mathrm{Var}(xy)(\mathrm{Var}(x)\,\mathrm{Var}(y) - \mathrm{Cov}(x, y)^2) \\
&\quad\quad - \mathrm{Var}(x)\,\mathrm{Cov}(xy, y)^2 - \mathrm{Var}(y)\,\mathrm{Cov}(xy, x)^2 \\
&\quad\quad + 2\,\mathrm{Cov}(x, y)\,\mathrm{Cov}(xy, y)\,\mathrm{Cov}(xy, x) \\
&= \Psi \ ,
\end{aligned}$$

so that dividing both sides by Ψ yields the desired result. □

Finally, we are ready to prove Lemma 4.9, which is the last result needed to establish that the variance estimator is unbiased. We restate the lemma below.

Lemma 4.9. *Fix a pair of outcome units $i \neq j \in V_o$. If $\Psi_{i,j} \neq 0$, then the covariance between individual treatment effect estimates $\hat{\tau}_i$ and $\hat{\tau}_j$ may be expressed as*

$$\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) = 4 \cdot \mathbb{E}[y_i(\boldsymbol{z}) y_j(\boldsymbol{z}) R_{i,j}] \ .$$

Proof. We may use the properties of the random variable $S_{i,j}$ proved in Lemma C.6 together with the linear response assumption to obtain

$$\begin{aligned}
\mathbb{E}[y_i(\boldsymbol{z}) y_j(\boldsymbol{z}) S_{i,j}] &= \mathbb{E}[(\beta_i x_i + \alpha_i)(\beta_j x_j + \alpha_j) S_{i,j}] \\
&= \beta_i \beta_j\, \mathbb{E}[x_i x_j S_{i,j}] + \alpha_i \alpha_j\, \mathbb{E}[S_{i,j}] + \beta_i \alpha_j\, \mathbb{E}[x_i S_{i,j}] + \beta_j \alpha_i\, \mathbb{E}[x_j S_{i,j}] \\
&= \beta_i \beta_j \ ,
\end{aligned}$$

where the first line follows from the linear-response assumption, the second line follows from expanding terms, and the third line follows from Lemma C.6. Next, we write the covariance between the individual treatment effect estimators. Because the individual

treatment effect estimators are unbiased,

$$\begin{aligned}
\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j) &= \mathbb{E}[\hat{\tau}_i\hat{\tau}_j] - \mathbb{E}[\hat{\tau}_i]\,\mathbb{E}[\hat{\tau}_j] \\
&= \mathbb{E}[\hat{\tau}_i\hat{\tau}_j] - \tau_i\tau_j \\
&= 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})y_j(\boldsymbol{z})\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big(\frac{x_j - \mathbb{E}[x_j]}{\mathrm{Var}(x_j)}\Big)\Big] - 4 \cdot \beta_i\beta_j \\
&= 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})y_j(\boldsymbol{z})\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big(\frac{x_j - \mathbb{E}[x_j]}{\mathrm{Var}(x_j)}\Big)\Big] - 4 \cdot \mathbb{E}[y_i(\boldsymbol{z})y_j(\boldsymbol{z})S_{i,j}] \\
&= 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})y_j(\boldsymbol{z})\Big[\Big(\frac{x_i - \mathbb{E}[x_i]}{\mathrm{Var}(x_i)}\Big)\Big(\frac{x_j - \mathbb{E}[x_j]}{\mathrm{Var}(x_j)}\Big) - S_{i,j}\Big]\Big] \\
&= 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})y_j(\boldsymbol{z})R_{i,j}\Big] \qquad \square
\end{aligned}$$

We are now ready to prove Theorem 4.10, which establishes unbiasedness of the variance estimator.

Theorem 4.10. *Under the conditions in Lemmas 4.8 and 4.9, the variance estimator of the ERL point estimator is unbiased, i.e.* $\mathbb{E}[\widehat{\mathrm{Var}}(\hat{\tau})] = \mathrm{Var}(\hat{\tau})$.

Proof. We may calculate the expectation of the variance estimate $\widehat{\mathrm{Var}}(\hat{\tau})$ as

$$\begin{aligned}
\mathbb{E}[\widehat{\mathrm{Var}}(\hat{\tau})] &= \mathbb{E}\left[\frac{4}{n^2}\sum_{i=1}^{n}\Big[y_i(\boldsymbol{z})^2 Q_i + \sum_{j\neq i} y_i(\boldsymbol{z})y_j(\boldsymbol{z})R_{i,j}\Big]\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left[4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})^2 Q_i\Big] + \sum_{j\neq i} 4 \cdot \mathbb{E}\Big[y_i(\boldsymbol{z})y_j(\boldsymbol{z})R_{i,j}\Big]\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\Big[\mathrm{Var}(\hat{\tau}_i) + \sum_{j\neq i}\mathrm{Cov}(\hat{\tau}_i, \hat{\tau}_j)\Big] \\
&= \mathrm{Var}(\hat{\tau}) \ ,
\end{aligned}$$

where the second equality follows by linearity of expectation and the third equality follows from Lemmas 4.8 and 4.9. $\square$

C.3 Exposure-Design and Correlation Clustering

In this section, we prove the relationship between Exposure-Design, its reformulation Corr-Clust, the previously proposed correlation clustering design of Pouget-Abadie et al. (2019), and other correlation clustering variants. A summary of the results are:

- In Section C.3.1, we show that the Exposure-Design may be reformulated as the clustering problem, Corr-Clust.

- In Section C.3.2, we compare Exposure-Design to the correlation clustering-based design presented in Pouget-Abadie et al. (2019). In particular, we prove that their design is equivalent to Exposure-Design when the trade-off parameter is set as $\phi = 1/(n-1)$ and no constraint is placed on cluster sizes, i.e. $k = m$.
- In Section C.3.3, we compare Corr-Clust to other correlation clustering variants. In particular, we prove that (unconstrained) Corr-Clust may be viewed as an instance of the weighted maximization correlation clustering considered by Charikar et al. (2005); Swamy (2004) but with a possibly large additive constant which prevents an approximation-preserving reduction.

To begin, we demonstrate how to re-write the Corr-Clust objective using matrix notation. Let $\omega_{i,j} \in \mathbb{R}$ be the weights for pairs $i, j \in [m]$ and let Ω be the m-by-m matrix whose (i,j)th entry is $\omega_{i,j}$. For a partition $\mathcal{C}$ of the indices $[m]$, let $Z_{\mathcal{C}}$ be the m-by-m matrix where the (i,j)th entry is 1 if i and j are in the same cluster of $\mathcal{C}$ and 0 otherwise. Then, we may express the Corr-Clust objective as

$$\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{i,j} [Z_{\mathcal{C}}]_{(i,j)} = \operatorname{tr}(\Omega \; Z_{\mathcal{C}}) \ .$$

Throughout the remainder of the section, it will be useful to write the Corr-Clust objective using this matrix notation.

C.3.1 Reformulating Exposure-Design as Corr-Clust

We are now ready to prove Proposition 4.12, which we restate here for completeness.

Proposition 4.12. *For each pair of diversion units $i, j \in V_d$, define the value $\omega_{i,j} \in \mathbb{R}$ as*

$$\omega_{i,j} = (1+\phi) \sum_{k=1}^{m} w_{k,i} w_{k,j} - \phi \Big(\sum_{k=1}^{m} w_{k,i} \Big) \Big(\sum_{k=1}^{m} w_{k,j} \Big) \ , \tag{4.2}$$

where $w_{k,i}$ is the weight of the edge between the kth outcome unit and the ith diversion unit. Exposure-Design *is equivalent to the following clustering problem:*

$$\max_{\text{clusterings } \mathcal{C}} \quad \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} \ . \tag{Corr-Clust}$$

Proof. Recall that the objective of Exposure-Design is defined as

$$\sum_{i=1}^{n} \operatorname{Var}(x_i) - \phi \sum_{i \neq j} \operatorname{Cov}(x_i, x_j) \ ,$$

where the expectation in the variance and covariance terms is taken with respect to the random assignment vector $\boldsymbol{z} \in \{\pm 1\}^m$, which is drawn from the cluster design

given by $\mathcal{C}$. Recall that the exposures are given by $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z}$. Using matrix notation, we can more compactly represent this objective as

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{Var}(x_i) - \phi \sum_{i \neq j} \mathrm{Cov}(x_i, x_j) &= \mathrm{tr}\Big(\big(\boldsymbol{I} - \phi(\mathbf{1}\mathbf{1}^\intercal - \boldsymbol{I}) \big) \, \mathrm{Cov}(\boldsymbol{x}) \Big) \\
&= \mathrm{tr}\Big(\big((1+\phi)\boldsymbol{I} - \phi \mathbf{1}\mathbf{1}^\intercal \big) \, \mathrm{Cov}(\boldsymbol{x}) \Big) \\
&= \mathrm{tr}\Big(\big((1+\phi)\boldsymbol{I} - \phi \mathbf{1}\mathbf{1}^\intercal \big) \, \mathrm{Cov}(\boldsymbol{W}\boldsymbol{z}) \Big) \\
&= \mathrm{tr}\Big(\big((1+\phi)\boldsymbol{I} - \phi \mathbf{1}\mathbf{1}^\intercal \big) \boldsymbol{W} \, \mathrm{Cov}(\boldsymbol{z}) \boldsymbol{W}^\intercal \Big) \\
&= \mathrm{tr}\Big(\boldsymbol{W}^\intercal \big((1+\phi)\boldsymbol{I} - \phi \mathbf{1}\mathbf{1}^\intercal \big) \boldsymbol{W} \, \mathrm{Cov}(\boldsymbol{z}) \Big)
\end{aligned}$$

where we have used properties of trace and covariance. Because $\boldsymbol{z}$ is drawn from an independent cluster design, the (i,j)th entry of the covariance matrix $\mathrm{Cov}(\boldsymbol{z})$ is 1 if diversion units i and j are in the same cluster and 0 otherwise. Thus, by the observation above, this clustering objective is a correlation clustering where the weights are given by the matrix

$$\Omega = \boldsymbol{W}^\intercal \big((1+\phi)\boldsymbol{I} - \phi \mathbf{1}\mathbf{1}^\intercal \big) \boldsymbol{W} \ .$$

By inspection, we have that the (i,j)th entry of this matrix Ω is

$$\omega_{i,j} = (1+\phi) \sum_{k=1}^{n} w_{k,i} w_{k,j} - \phi \Big(\sum_{k=1}^{n} w_{k,i} \Big) \Big(\sum_{k=1}^{n} w_{k,j} \Big) \ ,$$

as desired. □

C.3.2 An instance of Exposure-Design when $\phi = 1/(n-1)$

Now we demonstrate that the correlation clustering objective proposed in Pouget-Abadie et al. (2019) is a special case of Exposure-Design when $\phi = 1/(n-1)$ and no constraint is placed on cluster sizes, i.e. $k = m$. Before giving the formal statement, we re-introduce the clustering objective in that paper; that is,

$$\max_{\text{clusterings } \mathcal{C}} \mathbb{E}\left[\sum_{i=1}^{n} \left(x_i - \Big(\frac{1}{n} \sum_{j=1}^{n} x_j \Big) \right)^2 \right] \ , \qquad \text{(Exposure-Spread)}$$

where the expectation is with respect to the treatment vector $\boldsymbol{z} \in \{\pm 1\}^m$ drawn according to the independent cluster design given by $\mathcal{C}$. The quantity in the expectation is a measure of the spread of the exposures. We remark that in Pouget-Abadie et al. (2019), the exposures are called "doses" and the quantity in the expectation is referred to as the "empirical dose variance".

Proposition C.7. *Up to additive and multiplicative constants,* Exposure-Spread *is equivalent to* Exposure-Design *when the trade-off parameter is set to* $\phi = 1/(n-1)$.

Proof. Let us denote the exposure spread by

$$Q = \sum_{i=1}^{n} \left(x_i - \left(\frac{1}{n} \sum_{j=1}^{n} x_j \right) \right)^2 ,$$

Note that the exposure spread is equal to the ℓ_2 norm of the *de-meaned* exposure vector $\bar{\boldsymbol{x}} = (\bar{x_1}, \bar{x_2}, \dots \bar{x_n})$, where

$$\bar{x_i} = x_i - \left(\frac{1}{n} \sum_{j=1}^{n} x_j \right) .$$

The entire de-meaned exposure vector may be written as $\bar{\boldsymbol{x}} = (\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{x}$. Using the fact that this matrix is a projection and that the exposure vector is $\boldsymbol{x} = \boldsymbol{W}\boldsymbol{z}$, we can write the exposure spread as

$$\begin{aligned} Q = \|\bar{\boldsymbol{x}}\|^2 = \|(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{x}\|^2 &= \boldsymbol{x}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)^2\boldsymbol{x} \\ &= \boldsymbol{x}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{x} = \boldsymbol{z}^\intercal\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z} \ . \end{aligned}$$

Finally, the expectation of the exposure spread may be written as

$$\begin{aligned} \mathbb{E}[Q] &= \mathbb{E}\left[\boldsymbol{z}^\intercal\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z} \right] && \text{(from above)} \\ &= \mathbb{E}\left[\operatorname{tr}\Big(\boldsymbol{z}^\intercal\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z} \Big) \right] && \text{(trace of a scalar)} \\ &= \mathbb{E}\left[\operatorname{tr}\Big(\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\boldsymbol{z}\boldsymbol{z}^\intercal \Big) \right] && \text{(cyclic property of trace)} \\ &= \operatorname{tr}\Big(\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\,\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal] \Big) && \text{(linearity of trace)} \\ &= \operatorname{tr}\Big(\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\,\operatorname{Cov}(\boldsymbol{z}) \Big) + c \ , \end{aligned}$$

where the value c in the last line is $c = \operatorname{tr}\Big(\boldsymbol{W}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\,\mathbb{E}[\boldsymbol{z}]\,\mathbb{E}[\boldsymbol{z}]^\intercal \Big)$, which follows from $\operatorname{Cov}(\boldsymbol{z}) = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\intercal] - \mathbb{E}[\boldsymbol{z}]\,\mathbb{E}[\boldsymbol{z}]^\intercal$ and linearity of trace. Moreover, when the probability of treatment assignment p is fixed, this value c is a constant with respect to the clustering being chosen.

Observe that by setting $\phi = 1/(n-1)$ and multiplying by a factor $(n-1)/n$, the

EXPOSURE-DESIGN objective becomes

$$\frac{n-1}{n}\operatorname{tr}\Big(\boldsymbol{W}^\intercal\big((1+\frac{1}{n-1})\boldsymbol{I}-\frac{1}{n-1}\mathbf{1}\mathbf{1}^\intercal\big)\boldsymbol{W}\operatorname{Cov}(\boldsymbol{z})\Big)=\operatorname{tr}\Big(\boldsymbol{W}^\intercal(\boldsymbol{I}-\frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{W}\operatorname{Cov}(\boldsymbol{z})\Big)\ .$$

Thus, the EXPOSURE-SPREAD objective is equivalent (up to additive and multiplicative constants) to the EXPOSURE-DESIGN objective when $\phi = 1/(n-1)$. □

C.3.3 Comparison to other correlation clustering variants

Recall that we defined the objective of the correlation clustering variant CORR-CLUST as

$$\sum_{C_r\in\mathcal{C}}\sum_{i,j\in C_r}\omega_{i,j}\ ,$$

where $\omega_{i,j}$ is defined for each pair of diversion units $i, j \in V_d$ as

$$\omega_{i,j}=(1+\phi)\sum_{k=1}^{n}w_{k,i}w_{k,j}-\phi\Big(\sum_{k=1}^{n}w_{k,i}\Big)\Big(\sum_{k=1}^{n}w_{k,j}\Big)\ ,$$

and $w_{k,i}$ is the weight of the edge between the kth outcome unit and the ith diversion unit. Observe that the term $\omega_{i,j}$ can take positive or negative values.

The maximization weighted correlation clustering variant considered by Charikar et al. (2005); Swamy (2004) is defined as follows. Let $G = (V, E)$ be a graph where each edge $e = (i, j) \in E$ has two *non-negative* weights: $w_{in}(i, j)$ and $w_{out}(i, j)$. Given a clustering $\mathcal{C}$, an edge $e = (i, j)$ is said to be *in-cluster* if i and j are in the same cluster and *out-cluster* otherwise. The objective function for a given clustering is given by

$$\sum_{\substack{\text{in-cluster}\\ \text{edges } e}} w_{in}(e)+\sum_{\substack{\text{out-cluster}\\ \text{edges } e}} w_{out}(e) \qquad \text{(CORR-CLUST-CS)}$$

We now show that the CORR-CLUST objective may be written as an instance of the CORR-CLUST-CS objective, but with the addition of a large additive constant. Again, we stress that this reduction is primarily for aesthetic comparison purposes because the appearance of the large additive constant prevents any meaningful approximation-preserving reduction.

Proposition C.8. *Our formulation* CORR-CLUST *may be viewed as an instance of* CORR-CLUST-CS *with a large additive constant. More precisely, let* $w_{in}(i, j) = \max\{0, \omega_{i,j}\}$ *and* $w_{out}(i, j) = \min\{0, \omega_{i,j}\}$. *For a clustering* $\mathcal{C}$, *we have that the objectives are related by*

$$\sum_{C_r\in\mathcal{C}}\sum_{i,j\in C_r}\omega_{i,j}-\sum_{i=1}^{n}\sum_{j=1}^{n}\min\{0,\omega_{i,j}\}=\sum_{\substack{\textit{in-cluster}\\ \textit{edges } e}} w_{in}(e)+\sum_{\substack{\textit{out-cluster}\\ \textit{edges } e}} w_{out}(e)$$

Proof. For each pair of diversion units i, j, define $\omega^+_{i,j} = \max\{0, \omega_{i,j}\}$ and $\omega^-_{i,j} = -\min\{0, \omega_{i,j}\}$. Observe that $\omega_{i,j} = \omega^+_{i,j} + \omega^-_{i,j}$ and so we can re-distribute the following sum as

$$\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \left(\omega^+_{i,j} + \omega^-_{i,j} \right) = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} + \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^-_{i,j} \ .$$

Subtracting the (instance-dependent) constant $\sum_{i=1}^{n} \sum_{j=1}^{n} \min\{0, \omega_{i,j}\}$ from both sides and rearranging yields

$$\begin{aligned}
&\sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega_{i,j} - \sum_{i=1}^{n} \sum_{j=1}^{n} \min\{0, \omega_{i,j}\} \\
&\qquad = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} + \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^-_{i,j} - \sum_{i=1}^{n} \sum_{j=1}^{n} \min\{0, \omega_{i,j}\} \\
&\qquad = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} - \sum_{C_r \neq C'_r \in \mathcal{C}} \sum_{\substack{i \in C_r \\ j \in C'_r}} \omega^-_{i,j} \\
&\qquad = \sum_{C_r \in \mathcal{C}} \sum_{i,j \in C_r} \omega^+_{i,j} + \sum_{C_r \neq C'_r \in \mathcal{C}} \sum_{\substack{i \in C_r \\ j \in C'_r}} (-\omega^-_{i,j}) \\
&\qquad = \sum_{\substack{\text{in-cluster} \\ \text{edges } e}} w_{in}(e) + \sum_{\substack{\text{out-cluster} \\ \text{edges } e}} w_{out}(e) \ .
\end{aligned}$$

Finally, observe that for each pair (i, j), the values $w_{in}(i, j)$ and $w_{out}(i, j)$ are non-negative so that the final equation is a valid objective function for the Corr-Clust-CS formulation. □