

MIT Open Access Articles

*Balancing Covariates in Randomized Experiments
with the Gram–Schmidt Walk Design*

The MIT Faculty has made this article openly available. **Please share**
how this access benefits you. Your story matters.

Citation: Harshaw, C., Sävje, F., Spielman, D. A., & Zhang, P. (2024). Balancing Covariates in Randomized Experiments with the Gram–Schmidt Walk Design. *Journal of the American Statistical Association*, 119(548), 2934–2946.

As Published: <https://doi.org/10.1080/01621459.2023.2285474>

Publisher: Taylor & Francis

Persistent URL: <https://hdl.handle.net/1721.1/164296>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution





Balancing Covariates in Randomized Experiments with the Gram–Schmidt Walk Design

Christopher Harshaw, Fredrik Sävje, Daniel A. Spielman & Peng Zhang

To cite this article: Christopher Harshaw, Fredrik Sävje, Daniel A. Spielman & Peng Zhang (2024) Balancing Covariates in Randomized Experiments with the Gram–Schmidt Walk Design, *Journal of the American Statistical Association*, 119:548, 2934-2946, DOI: [10.1080/01621459.2023.2285474](https://doi.org/10.1080/01621459.2023.2285474)

To link to this article: <https://doi.org/10.1080/01621459.2023.2285474>



© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 09 Jan 2024.



[Submit your article to this journal](#)



Article views: 4993



[View related articles](#)



[View Crossmark data](#)



Citing articles: 11 [View citing articles](#)

Balancing Covariates in Randomized Experiments with the Gram–Schmidt Walk Design

Christopher Harshaw^a , Fredrik Sävje^b , Daniel A. Spielman^b, and Peng Zhang^c

^aMassachusetts Institute of Technology, Cambridge, MA; ^bYale University, New Haven, CT; ^cRutgers University, New Brunswick, NJ

ABSTRACT

The design of experiments involves a compromise between covariate balance and robustness. This article provides a formalization of this tradeoff and describes an experimental design that allows experimenters to navigate it. The design is specified by a robustness parameter that bounds the worst-case mean squared error of an estimator of the average treatment effect. Subject to the experimenter's desired level of robustness, the design aims to simultaneously balance all linear functions of potentially many covariates. Less robustness allows for more balance. We show that the mean squared error of the estimator is bounded in finite samples by the minimum of the loss function of an implicit ridge regression of the potential outcomes on the covariates. Asymptotically, the design perfectly balances all linear functions of a growing number of covariates with a diminishing reduction in robustness, effectively allowing experimenters to escape the compromise between balance and robustness in large samples. Finally, we describe conditions that ensure asymptotic normality and provide a conservative variance estimator, which facilitate the construction of asymptotically valid confidence intervals. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received February 2022
Accepted November 2023

KEYWORDS

Causal inference; Covariate balance; Treatment effects

1. Introduction

Randomized experiments are considered the most reliable way to estimate causal effects. Properly implemented randomization ensures that treatment effect estimators are unbiased. However, randomization does not ensure that estimators capture the true effect for any specific assignment of treatments. In an effort to make the estimators more precise, experimenters sometimes restrict the randomization to achieve covariate balance between treatment groups. A concern with this approach is that unobserved characteristics, including potential outcomes, may not be similar between the groups even if the observed characteristics are.

An idea that goes back to at least Efron (1971) is that the design of experiments involves a compromise between covariate balance and robustness. Randomization does not balance observed covariates to the same degree as a nonrandom assignment that specifically targets covariate balance, but randomization provides protection against imbalances on unobserved characteristics. Experimenters must weigh the robustness granted by randomness against possible gains in precision granted by balancing prognostically important covariates.



The first contribution of this article is a new formalization of the tradeoff between covariate balance and robustness. The formalization clarifies some ideas previously discussed by other authors and provides several new insights. We describe quantitative measures of both covariate balance and robustness, and we motivate the measures by showing that they characterize the


precision of the Horvitz–Thompson estimator of the average treatment effect. There is a fundamental tension between the two measures, as an experimenter cannot simultaneously achieve maximal robustness and fully balance covariates.

The second contribution of the article is the development of the Gram–Schmidt Walk design, which allows experimenters to navigate the tradeoff between balance and robustness. The design is specified by a parameter that bounds the worst-case mean squared error of the estimator. The design aims to simultaneously balance all linear functions of the covariates specified by the experimenter subject to meeting the worst-case guarantee. We describe several characterizations of the behavior of the design in finite samples. The main results are tight bounds on the mean squared error and on the tails of the sampling distribution of the treatment effect estimator.

We next investigate the asymptotic behavior of the estimator under the design. Under mild assumptions on the potential outcomes and the covariates, we show that the estimator is root- n consistent and that its limiting variance is the same as when all linear functions of the covariates are perfectly balanced. This means that the Gram–Schmidt Walk design allows experimenters to escape the balance–robustness tradeoff in large samples. The limiting variance of the estimator under the Gram–Schmidt Walk design is less than or equal to the limiting variance of other commonly used designs, such as rerandomization.

The final contribution of the article is to describe methods for inference. We provide a central limit theorem for the Horvitz–Thompson estimator under the Gram–Schmidt Walk

CONTACT Christopher Harshaw  charshaw@mit.edu  Massachusetts Institute of Technology, United States, Massachusetts, Cambridge.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

design, and provide a consistent, conservative estimator of the variance of the point estimator. Together, these results allow experimenters to construct conservative confidence intervals that are asymptotically valid.

A key discovery facilitating results in this article is a translation of the experimental design problem to a new type of problem in algorithmic discrepancy. A central problem of discrepancy theory is to partition a collection of vectors into two sets so that the sum of the vectors in each set is similar (Spencer 1985). This problem directly corresponds to finding a treatment assignment that maximizes covariate balance. However, algorithms for discrepancy minimization aim to produce a single partition, corresponding to a single assignment. Experimenters generally seek a distribution of assignments when they assign treatments, so as to achieve robustness from randomization. We argue that the experimental design problem is best interpreted as a *distributional* discrepancy problem. To tackle this problem, we take advantage of the Gram–Schmidt Walk algorithm of Bansal et al. (2019). This is a randomized algorithm, but the original authors used the randomization simply as a means to solve a non-distributional discrepancy problem. We leverage and deliberately amplify the randomized aspect to solve the distributional discrepancy problem. We also tighten and extend the analysis of the algorithm compared to Bansal et al. (2019) to be relevant for the experimental design problem. While we find the connection between these two fields insightful and important, an understanding of discrepancy theory is not required to understand the results in this article.

2. Related Work

The debate about the virtues of randomization goes back to the dawn of statistics. Student (1938) argued that randomization often is harmful because random assignments can only make treatment groups less comparable than what they would be under the most balanced assignment. This idea has more recently been discussed and extended by Bertsimas, Johnson, and Kallus (2015), Kasy (2016), Deaton and Cartwright (2018), and Kallus (2018). On the other hand, Fisher (1925, 1926) argued that randomization is desirable because it provides a certain level of robustness, in the form of unbiasedness, and facilitates well-motivated confidence intervals and testing. The first of Fisher’s points was extended by Wu (1981) to worst-case mean squared error, which is a more general robustness concept than unbiasedness (see also Kallus 2018; Basse, Ding, and Toulis 2022; Nordin and Schultzberg 2022; Bai 2023). Fisher’s second point is discussed and extended by Johansson, Rubin, and Schultzberg (2021).

A compromise between these two viewpoints is possible. While Wu (1981) demonstrates that there is no room to seek balance when robustness is our only objective, we might still be willing to accept a less robust design if it provides balance along dimensions we believe are important. This is the tradeoff between balance and robustness mentioned in the introduction. The idea can be traced back to Efron (1971), whose concept of “accidental bias” is closely related to our concept of robustness. This work has been extended by Kapelner et al. (2021) and a related idea based on a decision theoretical framework has been explored by Banerjee et al. (2020).

It is rare that experimenters assign treatment deterministically as suggested by Student (1938), but they do not necessarily assign treatments fully at random. Many designs fall in between the two extremes. Examples include the matched pair design (Greevy et al. 2004; Imai, King, and Nall 2009; Bruhn and McKenzie 2009), various stratified designs (Fisher 1935; Higgins, Sävje, and Sekhon 2016; Cytrynbaum 2021), and rerandomization (Lock Morgan and Rubin 2012; Li, Ding, and Rubin 2018). Existing analyses of these designs do not consider a formal balance–robustness tradeoff nor provide ways to navigate one.

To the best of our knowledge, there are only two prior designs that explicitly account for some version of the balance–robustness tradeoff. Krieger, Azriel, and Kapelner (2019) construct an algorithm that makes local changes to an assignment that is generated fully at random, aiming to produce a new assignment that is more balanced. They show that if there are few covariates, then few changes are needed to reach a highly balanced assignment, so the final assignment vector is similar to the one that was generated at random. Kapelner et al. (2022) investigate how to optimally select the acceptance criterion for rerandomization given a desired level of robustness.

3. Preliminaries

There are n units in the experiment, indexed by integers $[n] = \{1, \dots, n\}$. The experimenter randomly assigns a treatment $z_i \in \{\pm 1\}$ to each unit, and the assignments are collected in the random vector $\mathbf{z} = (z_1, \dots, z_n)$. We use $Z^+ = \{i \in [n] : z_i = 1\}$ and $Z^- = \{i \in [n] : z_i = -1\}$ to denote the random partition of the units into treatment and control groups. The *design* of the experiment is the distribution over the assignment vectors $\mathbf{z} \in \{\pm 1\}^n$.

Each unit has two potential outcomes: a_i , which is observed if $z_i = 1$, and b_i , which is observed if $z_i = -1$. We assume these potential outcomes are well-defined throughout the article, meaning that we rule out interference and other hidden versions of treatment. The observed outcome y_i for each unit is the random variable taking the value a_i when $z_i = 1$ and b_i when $z_i = -1$. It will prove convenient to collect the outcome variables into vectors:

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

Each unit has a vector of d covariates: $\mathbf{x}_i \in \mathbb{R}^d$. The largest covariate norm among the units is denoted $\xi = \max_{i \in [n]} \|\mathbf{x}_i\|$. The covariates are known to the experimenter prior to treatment assignment, so the experimental design may depend on them. The only randomness in the experiment comes from the assignment of treatment. The potential outcomes and covariates of the units are nonrandom and fixed, and we impose no assumptions on them at this point other than their existence.

The causal quantity of interest is the *average treatment effect*: $\tau = n^{-1} \sum_{i=1}^n (a_i - b_i)$. The average treatment effect cannot be directly observed, so it must be estimated. In this article, we restrict our attention to the Horvitz–Thompson estimator

$$\hat{\tau} = \frac{1}{n} \sum_{i \in Z^+} \frac{y_i}{\Pr(z_i = 1)} - \frac{1}{n} \sum_{i \in Z^-} \frac{y_i}{\Pr(z_i = -1)}.$$

This estimator is unbiased under designs that satisfy the positivity condition that the assignment probabilities are bounded away from zero and one for all units (Aronow and Middleton 2013). The aim of the experimenter when designing the experiment is to improve the precision of the estimator. To make the task concrete, we will primarily focus on the estimator's mean squared error, $\mathbb{E}[(\tau - \hat{\tau})^2]$, as our measure of precision.

For expositional purposes, we restrict our attention throughout the article to *symmetric designs* for which each unit is equally likely to receive either treatment: $\Pr(z_i = 1) = 1/2$. The extension of our results to settings with $\Pr(z_i = 1) \in (0, 1)$ is straightforward but notionally cumbersome, and it is therefore discussed in Section S8.1 in the supplement.

The error of the Horvitz–Thompson estimator for a particular assignment can be shown to depend on the potential outcomes only through their sum: $\boldsymbol{\mu} = \mathbf{a} + \mathbf{b}$. For short, we refer to $\boldsymbol{\mu}$ as the *potential outcome vector*. This insight allows us to derive the mean square error of the estimator under an arbitrary design.

Lemma 3.1. For all symmetric experimental designs, the mean squared error of the Horvitz–Thompson estimator is

$$\mathbb{E}[(\hat{\tau} - \tau)^2] = n^{-2} \boldsymbol{\mu}^T \text{Cov}(\mathbf{z}) \boldsymbol{\mu}.$$

The lemma demonstrates that the mean squared error is a quadratic form in the covariate matrix of the treatment assignment vector, $\text{Cov}(\mathbf{z})$, evaluated at the (unknown) potential outcome vector $\boldsymbol{\mu}$. Properties of the design that affect the mean squared error are therefore completely captured by $\text{Cov}(\mathbf{z})$. This is a central insight motivating our work in this article, informing both our interpretation of the experimental design problem as well as the proposed design. The insight has been used previously to inform investigations of experimental designs, and the characterization of the precision of the estimator in Lemma 3.1 is similar to those given by Efron (1971) and Kapelner et al. (2021).

4. The Balance–Robustness Tradeoff

4.1. A Measure of Robustness

Researchers use experiments because they provide credible causal inferences without need for strong assumptions. For example, under mild moment conditions on the potential outcomes and independent treatment assignment, the Horvitz–Thompson estimator is unbiased and converges to the average treatment effect at a root- n rate no matter what the potential outcomes might be. Experiments are in this sense robust. An important insight for what is to come is that all experiments are not equally robust. We use a worst-case concept to quantify robustness: an experiment is said to be robust if the estimator is sufficiently precise for all possible potential outcomes under its design. Building on the work of Efron (1971) and Kapelner et al. (2021), we show that the operator norm of $\text{Cov}(\mathbf{z})$ characterizes the worst-case performance of the design.

Lemma 4.1. For all symmetric experimental designs, the worst-case mean squared error over the set of all potential outcomes with bounded magnitude is

$$\max_{\boldsymbol{\mu} \in \text{PO}(M)} \mathbb{E}[(\tau - \hat{\tau})^2] = \frac{M}{n} \|\text{Cov}(\mathbf{z})\|,$$

where $\text{PO}(M) = \{\boldsymbol{\mu} \in \mathbb{R}^n : n^{-1} \|\boldsymbol{\mu}\|^2 \leq M\}$.

Lemma 4.1 shows that the operator norm $\|\text{Cov}(\mathbf{z})\|$ captures how robust a design is. The norm increases as the correlation between the assignments becomes stronger, so designs with greater correlation are less robust. An implication is that designs with no correlation are most robust, as captured by the following proposition.

Proposition 4.2. All symmetric experimental designs satisfy the inequality $\|\text{Cov}(\mathbf{z})\| \geq 1$, and equality holds for the Bernoulli design. Thus, the Bernoulli design is min–max optimal for potential outcomes with bounded average magnitude, $\text{PO}(M)$, for any M .

The Bernoulli design assigns treatments independently between units, so $\text{Cov}(\mathbf{z}) = \mathbf{I}$ and $\|\text{Cov}(\mathbf{z})\| = 1$. The operator norm cannot be made smaller than one because the diagonal entries of $\text{Cov}(\mathbf{z})$ are always one for symmetric designs, and the operator norm is at least the maximum entry. Thus, Proposition 4.2 shows that an experimenter who seeks to maximize robustness, when formalized in this way, should use the Bernoulli design.

The operator norm $\|\text{Cov}(\mathbf{z})\|$ can be seen as a unitless measure of robustness, in the sense that it measures the multiplicative increase in the worst-case mean squared error compared to the min–max design. For example, if some design has $\|\text{Cov}(\mathbf{z})\| = 2$, then its worst-case mean squared error is twice as large as the worst-case mean squared error under the min–max design. The largest possible value of $\|\text{Cov}(\mathbf{z})\|$ is n , achieved by a minimally random design that assigns some $\mathbf{z}' \in \{\pm 1\}^n$ with probability $1/2$, and otherwise its negation $-\mathbf{z}'$.

4.2. A Measure of Covariate Balance

A robust design ensures that the estimator is reasonably precise no matter what the potential outcomes might be. It is possible to make the estimator more precise if the experimenter has prior knowledge about the units and uses this knowledge when designing the experiment. The experimenter would then forgo some robustness to improve precision for certain potential outcomes. If the prior knowledge is in the form of pretreatment covariates that are known to be predictive of the potential outcomes, then precision is improved by using a design that ensures balance between the treatment groups with respect to those covariates.

We collect the units' covariate vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ as rows of an n -by- d matrix \mathbf{X} . It will prove convenient to use the maximum row norm $\xi = \max_{i \in [n]} \|\mathbf{x}_i\|$ as a measure the magnitude of the covariates \mathbf{X} . To introduce and illustrate our notion of covariate balance, assume for the moment that the covariates are perfectly linearly predictive of the outcomes, so there exists a function $\boldsymbol{\beta}$ such that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. This will not be necessary for any of our results, but it will be helpful to illustrate our concept of covariate balance in this section. Using Lemma 3.1, we can write the mean square error as

$$n^2 \mathbb{E}[(\hat{\tau} - \tau)^2] = \boldsymbol{\beta}^T \mathbf{X}^T \text{Cov}(\mathbf{z}) \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}^T \text{Cov}(\mathbf{X}^T \mathbf{z}) \boldsymbol{\beta}.$$

To make the estimator more precise in this setting, we should pick a design that makes $\beta^T \text{Cov}(X^T z) \beta$ small. However, even if we somehow knew that the covariates were perfectly predictive, we would generally not know the function β ; we must consider a set of possible functions. As above, we can use an operator norm bound for this purpose. For all linear functions $\beta \in \mathbb{R}^d$, we have

$$\beta^T \text{Cov}(X^T z) \beta \leq \|\text{Cov}(X^T z)\| \times \|\beta\|^2.$$

Holding the magnitude $\|\beta\|$ fixed, the operator norm $\|\text{Cov}(X^T z)\|$ provides a guarantee on the mean square error when the covariates are perfectly predictive. If the operator norm is small, we know that the mean square error is small no matter how the potential outcomes are related to the covariates. The bound is sharp, so conversely, if the operator norm is large, then we know that there exists a function β for which the mean square error is large compared to the magnitude $\|\beta\|$. Importantly, $\|\text{Cov}(X^T z)\|$ does not depend on the potential outcomes, so experimenters can target it when designing their experiments. For these reasons, we will use the operator norm $\|\text{Cov}(X^T z)\|$ as our measure of covariate balance. The dependence on $\|\beta\|$ in the bound is inescapable because the mean square error depends on the magnitude of the outcomes, and $\|\beta\|$ captures the relative scaling of the covariates and potential outcomes. Holding the relative scaling fixed, $\|\beta\|$ can be seen as a type of complexity measure of the function β .

In Section S5.2 in the supplement, we show that $\|\text{Cov}(X^T z)\|$ cannot be made smaller than ξ^2 without imposing additional restrictions or assumptions. A practically relevant upper bound on $\|\text{Cov}(X^T z)\|$ is given by the Bernoulli design, which achieves $\|\text{Cov}(X^T z)\| = \|X^T X\|$. Under the assumptions we use for our large sample analysis, $\xi^2 = \mathcal{O}(d \log(n))$ and $\|X^T X\| = \mathcal{O}(n)$, so the relevant interval for $\|\text{Cov}(X^T z)\|$ is $[d \log(n), n]$ up to constant factors.

4.3. The Tradeoff

A design that is maximally robust requires uncorrelated treatment assignments, but a design that achieves maximal covariate balance typically requires highly correlated assignments. It is therefore not possible to construct a design that achieves both maximum robustness and maximum covariate balance, in the sense that it is not possible to make the operator norms $\|\text{Cov}(z)\|$ and $\|\text{Cov}(X^T z)\|$ small simultaneously.

Proposition 4.3. If the largest singular value of the covariate matrix is larger than the maximum norm of the covariate vectors, $\sigma_{\max}(X) > \xi = \max_{i \in [n]} \|\mathbf{x}_i\|$, then there does not exist a design that simultaneously minimizes $\|\text{Cov}(z)\|$ and $\|\text{Cov}(X^T z)\|$.

The proposition captures a tension between $\|\text{Cov}(z)\|$ and $\|\text{Cov}(X^T z)\|$. We refer to this tension as the *balance-robustness tradeoff*. The condition of the proposition ensures that there is not a single covariate vector that dominates the balance properties of the experiment. We always have $\sigma_{\max}(X) \geq \xi$, so the condition rules out the edge case $\sigma_{\max}(X) = \xi$. As we will discuss in Section 7, the typical rate for $\sigma_{\max}(X)$ is \sqrt{n} , while the typical rate for ξ is $\sqrt{d \log(n)}$, meaning that $\sigma_{\max}(X)$ generally is much larger than ξ .

Because they cannot achieve both balance and robustness, experimenters must navigate the balance-robustness tradeoff when they design their experiments. To better understand what is at stake, consider when the covariates are only somewhat predictive, so covariate balance would be useful, but they are not perfectly predictive, so the analysis in the previous section does not apply. As above, let $\beta \in \mathbb{R}^d$ be some linear function. For the purpose of the current discussion, it is not important exactly what this function is—whatever it might be, it is unknown when the experiment is designed. Decompose the potential outcome vector into the linear function evaluated at the covariates, $X\beta$, and a residual term, $\epsilon = \mu - X\beta$, so that $\mu = X\beta + \epsilon$. If the covariates were perfectly predictive, then there exists a function β such that $\epsilon = \mathbf{0}$. However, we here consider when the covariates are only partially predictive, in which case $\epsilon \neq \mathbf{0}$ no matter the choice of β .

For an arbitrary function $\beta \in \mathbb{R}^d$, the mean square error decomposes as

$$\begin{aligned} n^2 \mathbb{E}[(\tau - \hat{\tau})^2] &= \beta^T \text{Cov}(X^T z) \beta + \epsilon^T \text{Cov}(z) \epsilon \\ &\quad + 2\beta^T \text{Cov}(X^T z, z) \epsilon. \end{aligned}$$

The first term of this expression, $\beta^T \text{Cov}(X^T z) \beta$, corresponds to the part of the potential outcomes that can be explained by the function β . Following the logic of the previous section, making $\|\text{Cov}(X^T z)\|$ small makes the first term small for any function β of fixed magnitude. The second term, $\epsilon^T \text{Cov}(z) \epsilon$, corresponds to what cannot be explained by the function β . If the covariates are completely unresponsive, then we cannot do better than $\beta = \mathbf{0}$, so $\mu = \epsilon$, and the second term corresponds exactly to what was studied in Section 4.1. Hence, making $\|\text{Cov}(z)\|$ small, makes the second term small.

The third term of the decomposition is a cross term: $2\beta^T \text{Cov}(X^T z, z) \epsilon$. It is possible to characterize this cross term by considering properties of the matrix $\text{Cov}(X^T z, z)$. Indeed, we do this for the design described in this article, the Gram-Schmidt Walk design. However, for the purpose of illustrating the key tension in the balance-robustness tradeoff, such a detailed characterization would be a distraction. Instead, to understand the tradeoff, note that the cross term is bounded by the first two terms:

$$2\beta^T \text{Cov}(X^T z, z) \epsilon \leq \beta^T \text{Cov}(X^T z) \beta + \epsilon^T \text{Cov}(z) \epsilon.$$

Therefore, controlling the first two terms indirectly controls the cross-term; if the first two terms are small, so is the cross term.

When the covariates are only partially predictive of the potential outcomes, the decomposition tells us that the mean square error is determined by both $\|\text{Cov}(z)\|$ and $\|\text{Cov}(X^T z)\|$. Ideally, we would want to make both operator norms small, but this is not possible. Because of the balance-robustness tradeoff, we can generally make one of them small only by accepting that the other becomes larger. Experimenters therefore have to choose between a design that balances the covariates well or a design that is highly robust, or something in-between.

4.4. Balance-Robustness Design Guarantees

While it is not possible to make $\|\text{Cov}(z)\|$ and $\|\text{Cov}(X^T z)\|$ simultaneously small, it is possible to make them simultaneously

large. That is, there are designs that provide both poor covariate balance and poor robustness. We seek to avoid such designs.

Definition 4.4. An experimental design is said to provide a balance–robustness guarantee of (γ_z, γ_x) if it ensures that $\|\text{Cov}(\mathbf{z})\| \leq \gamma_z$ and $\|\text{Cov}(\mathbf{X}^T \mathbf{z})\| \leq \gamma_x$.

A design that provides a guarantee that both γ_z and γ_x are reasonably small navigates the balance–robustness tradeoff well. While it is not possible to attain the minimums of γ_z and γ_x simultaneously, we can consider the minimal pairs (γ_z, γ_x) . That is, a design that provides maximum covariate balance for a given level of robustness, or maximum robustness for a given level of covariate balance. The set of all such designs constitutes a Pareto frontier of the balance–robustness tradeoff. We argue that experimenters should, if possible, use designs that are on or close to this Pareto frontier.

Designs that provide a balance–robustness guarantee also implicitly yields a guarantee on the mean square error of the treatment effect estimator. Therefore, a design that better navigates the balance–robustness tradeoff, in the sense of being closer to the Pareto frontier, provides a sharper guarantee on the mean square error.

Theorem 4.5. For any symmetric experimental design with balance–robustness guarantee (γ_z, γ_x) , the mean squared error of the Horvitz–Thompson estimator is bounded as

$$n\mathbb{E}[(\hat{\tau} - \tau)^2] \leq \min_{\beta \in \mathbb{R}^d} \left[\frac{\gamma_z}{n} \|\mu - \mathbf{X}\beta\|^2 + \frac{\gamma_x}{n} \|\beta\|^2 + \frac{2\sqrt{\gamma_z \gamma_x}}{n} \|\mu - \mathbf{X}\beta\| \|\beta\| \right].$$

The theorem is a generalization of the characterizations of the mean square error in Sections 4.1 and 4.2 to settings with partially predictive covariates. The first term captures how well a linear function β predicts the potential outcomes using the covariates. This term can be made small if the potential outcome vector μ is close to the span of the covariates. The second term captures the magnitude of the function β . This term can be made small by using a function of small magnitude, typically meaning that the function does not predict the potential outcomes well. The third term is the cross term discussed in the previous section. The balance–robustness guarantee (γ_z, γ_x) determines the tradeoff between the terms, assigning more focus to either finding a function that predicts the outcomes well or one that is of small magnitude. If the covariates are predictive, in the sense that there exists a function $\beta \in \mathbb{R}^d$ such that the norm of $\epsilon = \mu - \mathbf{X}\beta$ is small, then making γ_x small will be more beneficial than making γ_z small. However, if no such function exists, so that the minimum of $\|\epsilon\|$ is approximately the same as $\|\mu\|$, then making γ_x small could cause harm by making γ_z large. The magnitude of the cross term is the geometric mean of the two leading terms, so if either of those terms are small, so will the cross term be.

The bound in Theorem 4.5 is tight, in the sense that it holds with equality for some potential outcomes and covariates, but there are situations in which the bound is quite loose. It is not the purpose of the theorem to give an exact characterization of the mean square error. Lemma 3.1 gives an exact characterization,

but it depends on the full covariance matrix of the assignment, so it is considerably more complex than Theorem 4.5. The purpose of the theorem is instead to show that the balance–robustness tradeoff, as formalized by the operator norms $\|\text{Cov}(\mathbf{z})\|$ and $\|\text{Cov}(\mathbf{X}^T \mathbf{z})\|$, is widely applicable and relevant. Crucially, the operator norms and the balance–robustness guarantee do not depend on the potential outcomes, so they can be used during the design stage of the experiment, before observing any potential outcomes. A sharper characterization of the mean square error would need to consider more intricate aspects of the design and how they interact with the potential outcomes, making the characterization less useful for the purpose of designing experiments, because experimenters generally do not have access to such information at the design stage.

5. The Gram–Schmidt Walk Design

The Gram–Schmidt Walk design is constructed to navigate the balance–robustness tradeoff. It is parameterized by $\phi \in [0, 1]$, which controls its balance–robustness guarantee.

A central aspect of the design is the construction of an augmented covariate vector $\mathbf{b}_i \in \mathbb{R}^{n+d}$ for each unit. This is a scaled concatenation of the unit’s raw covariate vector and a unit-unique indicator variable:

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

where $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)$ is the i th standard basis vector of dimension n and $\xi = \max_{i \in [n]} \|\mathbf{x}_i\|$ is the maximum covariate norm. We collect the augmented vectors $\mathbf{b}_1, \dots, \mathbf{b}_n$ as columns of an $(n + d)$ -by- n matrix \mathbf{B} .

The design uses the augmented covariate vectors as input to a slight modification of the Gram–Schmidt Walk algorithm of Bansal et al. (2019). This algorithm produces a random assignment vector $\mathbf{z} \in \{\pm 1\}^n$ with the property that the (random) difference between the within-group sums of the augmented vectors concentrates around zero with high probability. That is, $\mathbf{B}\mathbf{z} = \sum_{i \in Z^+} \mathbf{b}_i - \sum_{i \in Z^-} \mathbf{b}_i \approx \mathbf{0}$. By balancing the augmented covariate vectors, the Gram–Schmidt Walk design balances both the original raw covariate vectors and the unit-unique basis vectors \mathbf{e}_i .

The 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. The basis vectors are best balanced by assigning treatment fully at random, so the Gram–Schmidt Walk design induces less correlation between treatments when augmented covariate vectors mostly resemble the basis vectors. This is the way the design navigates the balance–robustness tradeoff. When $\phi = 1$, the augmented covariate vectors are exactly the orthogonal basis vectors. In that case, the Gram–Schmidt Walk design recovers the Bernoulli design.

The algorithm for sampling from the Gram–Schmidt Walk is described in Algorithm 1. It builds on a relaxation of 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 \mathbf{z}_t until it takes values in $\{\pm 1\}$. The initial fractional assignments are zero: $\mathbf{z}_1 = \mathbf{0}$. This

Algorithm 1: The Gram–Schmidt Walk

```

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

$$\mathbf{u}_t \leftarrow \arg \min_{\mathbf{u} \in \mathbb{R}^n} \|\mathbf{B}\mathbf{u}\|^2$$

   subject to  $u(p) = 1$ , and  $u(i) = 0$  for all  $i \notin \mathcal{A}$ 
8   Set  $\delta^+ \leftarrow |\max \Delta|$  and  $\delta^- \leftarrow |\min \Delta|$  where
 $\Delta = \{\delta \in \mathbb{R} : \mathbf{z}_t + \delta \mathbf{u}_t \in [-1, 1]^n\}$ .
9   Select a step size at random according to

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

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

```

means that the augmented covariate vectors start out perfectly balanced, because $\mathbf{B}\mathbf{z}_1 = \mathbf{B}\mathbf{0} = \mathbf{0}$. However, the initial assignments are not acceptable, because they are not in $\{\pm 1\}^n$. As the algorithm updates the fractional assignments, the fundamental tension is between maintaining good balance, as measured by $\mathbf{B}\mathbf{z}_t$, and making the assignments integral. 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 toward integrality.

A general implementation of the Gram–Schmidt Walk algorithm that explicitly constructs and solves the system of linear equations from scratch at each iteration would run in $\mathcal{O}(n^4 + n^3d)$ time. However, the structure of the augmented covariates allows us to construct a customized implementation that maintains a Cholesky factorization between iterations, improving the run time to $\mathcal{O}(n^2d)$. Section S6 in the supplement describes this implementation and proves its computational properties.

There are similarities between the Gram–Schmidt Walk design and the Cube Method of Deville and Tillé (2004), which is used in survey sampling. Both methods build on the idea that assignment vectors can be represented as vertices of a hypercube and that an assignment can be obtained through a random walk inside that hypercube. Indeed, many discrepancy minimization algorithms are based on such geometric interpretations. The Cube Method can be seen as a randomized version of an algorithm by Beck and Fiala (1981) for discrepancy minimization, followed by a rounding procedure. To the best of our knowledge, this connection has gone unnoticed by both the survey sampling and theoretical computer science communities. Unlike the Cube Method, which has two distinct phases, the

iterations of the Gram–Schmidt Walk design all take a similar form. The two-phase structure prevents the Cube Method from achieving balance–robustness guarantees comparable to those of the Gram–Schmidt Walk design, as the first phase does not consider how its updates affect the second phase.

6. Finite-Sample Properties

6.1. Martingale and Unbiasedness

A central property of the Gram–Schmidt Walk design is that the sequence of the fractional assignment vectors forms a martingale. This implies that the expectation of the assignments sampled from the design is zero, $\mathbb{E}[\mathbf{z}] = \mathbf{z}_1 = \mathbf{0}$, which in turn ensures unbiasedness of the Horvitz–Thompson estimator for the average treatment effect. These insights are formalized in the following lemma and corollary.

Lemma 6.1. The sequence of fractional assignments $\mathbf{z}_1, \mathbf{z}_2, \dots$ forms a martingale.

Corollary 6.2. Under the Gram–Schmidt Walk design, $\Pr(\mathbf{z}_i = \mathbf{1}) = 1/2$ for all $i \in [n]$. Thus, the Horvitz–Thompson estimator is unbiased under the design.

The relation $\mathbb{E}[\mathbf{z}] = \mathbf{z}_1$ holds for any initial fractional assignments, which provides control over the first moment of the assignment vector. We use this fact to extend the design to nonuniform assignment probabilities in Section S8.1 in the supplement.

6.2. Navigating the Tradeoff

The Gram–Schmidt Walk design is able to navigate the balance–robustness tradeoff because it balances the augmented covariate vectors well, as described in the following theorem. The proof, which is provided in the supplement, interprets the algorithm as implicitly constructing a random basis for the column space of \mathbf{B} , which reveals the connection between the Gram–Schmidt Walk and its namesake, the Gram–Schmidt orthogonalization procedure.

Theorem 6.3. Under the Gram–Schmidt Walk design, the covariance matrix of the vector of imbalances for the augmented covariates $\mathbf{B}\mathbf{z}$ is bounded as $\text{Cov}(\mathbf{B}\mathbf{z}) \preceq \mathbf{P}$, where $\mathbf{P} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ is the orthogonal projection onto the subspace spanned by the columns of \mathbf{B} .

The covariance matrix $\text{Cov}(\mathbf{B}\mathbf{z})$ in Theorem 6.3 captures how balanced the augmented covariates are. The theorem states that the augmented covariates are balanced because the projection matrix \mathbf{P} in the upper bound is small by construction: it has at most n eigenvalues that are one and d eigenvalues that are zero. With this result in hand, we are ready to investigate what balance–robustness guarantee the design provides.

Theorem 6.4. The Gram–Schmidt Walk design with parameter $\phi \in [0, 1]$ provides the balance–robustness guarantee

$$\gamma_z = \frac{1}{\phi} \quad \text{and} \quad \gamma_x = \frac{\xi^2}{1 - \phi}.$$

When $\phi = 1$, the Gram–Schmidt Walk design places all emphasis on robustness and the min–max optimal robustness guarantee of $\gamma_z = 1$ is obtained. When $\phi = 0$, all emphasis is instead placed on covariate balance and the balance guarantee $\gamma_x = \xi^2$ is obtained. As we noted in Section 4.2, γ_x cannot be made smaller than ξ^2 unless restrictions are imposed on the covariates. Intermediate values of the design parameter, $\phi \in (0, 1)$, interpolate between these two extremes. In this way, the design navigates the balance–robustness tradeoff, and it lets experimenters select a guarantee that is appropriate for their applications.

The balance–robustness guarantee (γ_z, γ_x) in Theorem 6.4 can be loose relative to the actual performance of the Gram–Schmidt Walk design, especially for values of ϕ near 0 and 1. For example, the Gram–Schmidt Walk design with $\phi = 1$ is exactly the Bernoulli design. In this case, we know that $\|\text{Cov}(\mathbf{X}^T \mathbf{z})\| = \|\mathbf{X}^T \mathbf{X}\|$, but Theorem 6.4 gives the vacuous bound $\|\text{Cov}(\mathbf{X}^T \mathbf{z})\| \leq \infty$. Similarly, when $\phi = 0$, we know that $\|\text{Cov}(\mathbf{z})\| \leq n$, but Theorem 6.4 again gives the vacuous bound of $\|\text{Cov}(\mathbf{z})\| \leq \infty$. This is largely a consequence of Theorem 6.4 giving a balance–robustness guarantee for arbitrary covariates. We would need to consider specific covariates to provide a sharper guarantee, and that would lead to a more complex bound. The purpose of Theorem 6.4 is to provide a finite-sample guarantee that is easy to understand and work with in practice. The Gram–Schmidt Walk design can sometimes perform considerably better than this guarantee.

6.3. Mean Squared Error

Theorems 4.5 and 6.4 together provide a bound on the mean square error of the Horvitz–Thompson estimator under the Gram–Schmidt Walk design. We use our understanding of the design to sharpen this bound, as described by the following theorem.

Theorem 6.5. The mean squared error of the Horvitz–Thompson estimator under the Gram–Schmidt Walk design is at most the minimum of the loss function of an implicit ridge regression of the sum of the potential outcome vectors $\boldsymbol{\mu} = \mathbf{a} + \mathbf{b}$ on the covariates:

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

The bound in Theorem 6.5 is the same as the bound in Theorem 4.5, but with the cross-term removed. Theorem 6.3 allows us to address the cross term directly, rather than use the cruder bound we used in Section 4. A cosmetic difference between Theorems 4.5 and 6.5 is that we here have written the bound in terms of the design parameter ϕ rather than the balance–robustness guarantee (γ_z, γ_x) .

The right-hand side of the bound in Theorem 6.5 is the scaled minimum loss of a ridge regression of the potential outcomes on the covariates. The design parameter ϕ determines the regularization penalty of the regression, giving more weight to either functions $\boldsymbol{\beta}$ that explain the potential outcomes well or functions of small magnitude. This is a manifestation of the balance–robustness tradeoff.

While the mean square error is bounded by the loss of a ridge regression, no regression is ever run. The estimator is the ordinary, unadjusted Horvitz–Thompson estimator. Indeed, the regression can never be run, because it involves all potential outcomes, and we only observe half of them. Theorem 6.5 instead highlights that the design assigns treatments in a way that makes the Horvitz–Thompson estimator behave as if such a regression had been run.

We can use Theorem 6.5 to characterize when it is beneficial to deviate from the mini-max design and set $\phi < 1$. We already know from the balance–robustness tradeoff that $\phi = 1$ is optimal when the covariates are completely unpredictable of the outcomes, but the tradeoff by itself does not tell us how predictive the covariates must be to make it useful to seek some covariate balance. In Section S9.8 in the supplement, we show that it is almost always beneficial to seek at least some covariate balance by setting $\phi < 1$ when using the Gram–Schmidt Walk design. One exception is small experiments with nearly unpredictable covariates.

6.4. Tail Behavior

Our characterization of the mean square error in the previous subsection gives only a limited view of the behavior of the Gram–Schmidt Walk design. To paint a more complete picture, we provide finite-sample valid tail bounds on the discrepancy of the augmented covariates, \mathbf{Bz} , and the Horvitz–Thompson estimator.

Bansal et al. (2019) used the martingale inequality of Freedman (1975) to show that the Gram–Schmidt Walk algorithm ensures that \mathbf{Bz} is a sub-Gaussian random vector with variance parameter $\sigma^2 \leq 40$. However, tail bounds based on $\sigma^2 = 40$ will generally be too loose to be informative and useful in a statistical context. The following theorem strengthens the analysis to variance parameter $\sigma^2 = 1$, which is tight. To achieve this result, we develop a new proof technique for establishing martingale concentration, which might be of independent interest. The proof technique is described in the supplement.

Theorem 6.6. Under the Gram–Schmidt Walk design, the vector \mathbf{Bz} is sub-Gaussian with variance parameter $\sigma^2 = 1$. That is, $\mathbb{E}[\exp(\langle \mathbf{Bz}, \mathbf{v} \rangle)] \leq \exp(\|\mathbf{v}\|^2 / 2)$ for all $\mathbf{v} \in \mathbb{R}^{n+d}$.

The proof appears in Section S3.5 in the supplement, and is based on a bound on the conditional expectation of an exponential quantity during a pivot phase. Bansal et al. (2019) bound this quantity using a lossy Taylor series approximations. In contrast, we analyze it directly.

Theorem 6.6 demonstrates 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. If we in Theorem 6.6 set $\mathbf{v} = n^{-1} \mathbf{B} (\mathbf{B}^T \mathbf{B})^{-1} \boldsymbol{\mu}$, we get $\langle \mathbf{Bz}, \mathbf{v} \rangle = \widehat{\tau} - \tau$. This allows us to use the theorem to also derive a finite-sample tail bound for the Horvitz–Thompson estimator itself.

Corollary 6.7. Under the Gram–Schmidt Walk design, the tails of the sampling distribution of the Horvitz–Thompson estimator are bounded in finite samples such that, for all $\gamma > 0$,

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

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

7. Large-Sample Properties

7.1. Asymptotic Regime and Assumptions

Following convention in the design-based causal inference literature, we consider a sequence of finite populations of growing size in our large sample analysis. All aspects of the experiment, including the potential outcomes and the design parameter, are thus indexed by n . However, we leave the indexing implicit for notational clarity. Our analysis focuses on the limiting behavior of the estimator and the design under conditions on the sequence of experiments.

Assumption 7.1 (Outcome regularity). The fifth moments of the potential outcomes are asymptotically bounded: $n^{-1} \|\mathbf{a}\|_5^5 = \mathcal{O}(1)$ and $n^{-1} \|\mathbf{b}\|_5^5 = \mathcal{O}(1)$.

Assumption 7.2 (Covariate regularity). The singular values of the covariate matrix are asymptotically bounded as $\sigma_{\min}(\mathbf{X}) = \Omega(n^{1/2})$ and $\sigma_{\max}(\mathbf{X}) = \mathcal{O}(n^{1/2})$.

Assumption 7.3 (No extreme outliers). The maximum squared norm of the covariate vectors grows at most at the rate $\xi^2 = \max_{i \in [n]} \|\mathbf{x}_i\|^2 = \mathcal{O}(d \log(n))$.

Assumption 7.4 (Covariate dimensions). The number of covariates grows at most at the rate $d = \mathcal{O}(n^{1/10-\varepsilon})$ for some $\varepsilon > 0$.

Outcome regularity (Assumption 7.1) ensures that there are no extreme outliers among the potential outcomes. The lower bound on the smallest singular value in Assumption 7.2 ensures that the moment matrix of the covariates, $n^{-1} \mathbf{X}^T \mathbf{X}$, is invertible. Together with the upper bound on the largest singular value, this ensures that the matrix is well-conditioned. No extreme outliers (Assumption 7.3) ensures that the magnitude of the largest covariate vector \mathbf{x}_i is not too large, and Assumption 7.4 ensures that there are not too many covariates relative to the number of units. The last three assumptions concern the covariates, which are observed at the design stage. Experimenters can therefore calculate and inspect the quantities in the assumptions before committing to a design. They can also transform the covariates, for example by deleting columns, so as to better satisfy the assumptions. If the covariates were to be drawn at random from some large population, Assumptions 7.2 and 7.3 would be satisfied with a probability approaching one if, for example, the population distribution of the covariate vector was sub-Gaussian and the population moment matrix was invertible.

In the supplement, we show that some of these assumptions can be made more general or otherwise relaxed, at the cost of added complexity of the theorems and proofs. For example, if Assumption 7.1 is strengthened to uniformly bounded outcomes, then Assumption 7.4 can be weakened to $d = \mathcal{O}(n^{1/6-\varepsilon})$ for some $\varepsilon > 0$.

7.2. Consistency

The Horvitz–Thompson estimator is root- n consistent under most designs, including the Bernoulli design, and we want to ensure that the Gram–Schmidt Walk design does not lead to a slower rate of convergence.

Theorem 7.5. Suppose that outcome regularity holds (Assumption 7.1) and that the design parameter is asymptotically bounded away from zero, $\phi = \Omega(1)$. Then, the Horvitz–Thompson estimator under the Gram–Schmidt Walk design is root- n consistent for the average treatment effect: $\widehat{\tau} - \tau = \mathcal{O}_p(n^{-1/2})$.

The theorem shows that the Gram–Schmidt Walk design achieves root- n consistency as long as experimenters do not let the design parameter approach zero, assigning at least some weight to robustness in the design tradeoff. The theorem uses bounded fifth outcome moments as stipulated by Assumption 7.1, but the proof, which appears in the supplement, makes clear that this can be relaxed to bounded second moments. This is the same condition required for root- n consistency under conventional designs. Indeed, if the second moments are not bounded, and no other assumptions are imposed on the outcomes, then there exists no design that is root- n consistent.

7.3. Limiting Variance

The limiting variance of the estimator under the Gram–Schmidt Walk design depends on the sequence of the design parameter ϕ . All else equal, it is easier to achieve a certain level of covariate balance when the sample is larger. By letting ϕ approach one as the sample grows, it is possible to approach a setting with both maximal covariate balance and maximal robustness, effectively escaping the balance–robustness tradeoff in large samples. The following theorem formalizes this insight.

Theorem 7.6. Suppose outcome and covariate regularity holds (Assumptions 7.1 and 7.2). Further suppose that the design parameter approaches one at a sufficiently slow rate, so that $1 - \phi = o(1)$ and $1 - \phi = \omega(\xi^2/n)$. Then, a tight asymptotic upper bound on the normalized variance of the Horvitz–Thompson estimator under the Gram–Schmidt Walk design is

$$\limsup_{n \rightarrow \infty} [n \text{Var}(\widehat{\tau}) - V_{\text{GSW}}] \leq 0,$$

where $V_{\text{GSW}} = n^{-1} \min_{\beta} \|\mu - X\beta\|^2$ is the mean square residuals from a best least squares linear approximation of the potential outcomes using the covariates.

The theorem describes the precision of the estimator in large samples when the design parameter approaches one. It is an upper bound because V_{GSW} does not fully characterize the behavior of the design in the orthogonal complement of the covariate space. While we have not found any sequences of potential outcomes for which the design performs better than the bound, we have not shown that none exist. However, the bound is instance tight, in the sense that there always exist sequences of potential outcomes such that it holds with equality, no matter what the covariates might be. We conjecture that the

bound characterizes the asymptotic variance for most potential outcomes, in the sense that the bound holds with equality for all potential outcomes under mild regularity conditions. The motivation for this conjecture is that most eigenvalues of $\text{Cov}(\mathbf{z})$ will approach one when $\phi \rightarrow 1$.

Note that $V_{\text{GSW}} = n^{-1} \min_{\beta} \|\mu - \mathbf{X}\beta\|^2$ would be attainable as the variance in finite samples if we somehow had access to all potential outcomes, so we could calculate the best linear approximation, and then use the residuals from this regression as outcomes in the experiment. This procedure is of course infeasible, because we never observe all potential outcomes. Nevertheless, V_{GSW} marks the lowest variance achievable by balancing linear functions. [Theorem 7.6](#) shows that we can attain this lower limit asymptotically using the Gram–Schmidt Walk design. Another way to achieve V_{GSW} as the limiting variance is to do covariate adjustment in the estimation stage, as described by [Lin \(2013\)](#). However, such ex post covariate adjustment is not well-understood in finite samples, and introduces the risk of specification searching, so called p-hacking. The Gram–Schmidt Walk design achieves V_{GSW} as the limiting variance by design, using the unadjusted Horvitz–Thompson estimator in the estimation stage.

It is important to let the design parameter approach one, $\phi \rightarrow 1$, but never set it exactly to one. If we were to set $\phi = 1$, we would get the Bernoulli design. The normalized variance would then be $\|\mu\|^2/n$, which can be considerably larger than V_{GSW} . [Theorem 7.6](#) also requires that ϕ approaches one at a sufficiently slow rate, so that $n(1 - \phi)/\xi^2 \rightarrow \infty$. Given that ξ^2 typically will be of considerably lower order than n , this rate condition is quite forgiving. That is, the theorem describes the asymptotic behavior of the estimator for a wide range of sequences of ϕ , and experimenters have substantial latitude in selecting the design parameter. For example, if [Assumption 7.3](#) holds, so $\xi^2 = \mathcal{O}(d \log(n))$, and $d = o(n^\alpha)$ for some $0 < \alpha < 1$, then setting $\phi = 1 - Cn^{\alpha-1} \log(n)$, for some constant $C \in \mathbb{R}^+$, ensures that the rate condition holds. Note, however, that [Theorem 7.6](#) does not require [Assumption 7.3](#) to hold. Note also that ξ^2 and n are known by the experimenter at the design stage, so they can select ϕ to ensure that the rate condition holds.

In [Section S4.4](#) in the supplement, we analyze the limiting variance when ϕ is fixed asymptotically, relaxing the condition that the design parameter approaches one. [Chatterjee, Dey, and Goswami \(2023\)](#) provide an improved analysis of the limiting variance under the Gram–Schmidt Walk design when ϕ is fixed asymptotically under slightly different assumptions than the ones we use.

7.4. Asymptotic Normality

The finite-sample tail bounds for the Horvitz–Thompson estimator described in [Section 6.4](#) will often be loose in large samples. The following theorem describes when the distribution of the estimator approaches a normal distribution as the sample grows.

Theorem 7.7. Suppose that [Assumptions 7.1, 7.2, 7.3](#), and [7.4](#) hold. Further suppose that the limiting distribution of the estimator is nondegenerate, in the sense that $n\text{Var}(\widehat{\tau}) = \Omega(1)$. Then, if the design parameter is asymptotically bounded away

from zero, $\phi = \Omega(1)$, the limiting distribution of the Horvitz–Thompson estimator under the Gram–Schmidt Walk design is the standard normal distribution:

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

We require that $n\text{Var}(\widehat{\tau}) = \Omega(1)$ to avoid situations in which the estimator converges faster than the parametric rate. There are sequences of experiments that satisfy our conditions for which $n\text{Var}(\widehat{\tau}) \rightarrow 0$, but they are all knife-edge cases that are of little practical relevance, so non-degeneracy can be seen as a regularity condition. The non-degeneracy condition has been used previously in the design-based causal inference literature; examples include [Condition 6](#) in [Aronow and Samii \(2017\)](#) and [Assumption 5](#) in [Leung \(2022\)](#).

To the best of our knowledge, the technique we use to prove [Theorem 7.7](#) has not previously been used in the design-based causal inference literature. Central limit theorems build on the insight that an appropriately scaled sum of sufficiently many weakly dependent random variables tend to be close to a normal distribution. The conventional proof strategy in this setting is to analyze the terms of a linear estimator, which in our case would be $2y_i z_i/n$. Instead of following this convention, we reinterpret the estimator as being the sum of the updates of the assignments (to all units) in each iteration of the Gram–Schmidt Walk algorithm. That is, using the notation from [Section 5](#), we interpret the estimator to be the sum of terms of the form $\delta_t \mathbf{u}_t^T \mu/n$ over the iterations $t \in [T]$. This allows us to form a martingale difference sequence for $\widehat{\tau} - \tau$, to which we can apply the martingale central limit theorem by [McLeish \(1974\)](#). The proof appears in the supplement.

8. Inference

8.1. Variance Bound and Estimator

The first step toward constructing our confidence intervals is to estimate the variance of the estimator under the Gram–Schmidt Walk design. However, the variance depends on joint features of the two potential outcomes, which are inherently unobservable, so it is not directly estimable. This is a common problem in design-based causal inference. We follow the conventional solution of estimating an upper bound for the variance, which acts as a conservative estimator. The bound we use is based on the following decomposition of the limiting variance.

Proposition 8.1. The limiting variance of the Horvitz–Thompson estimator under the Gram–Schmidt Walk design can be written

$$\begin{aligned} nV_{\text{GSW}} = \min_{\beta} \|\mu - \mathbf{X}\beta\|^2 &= 2 \min_{\beta} \|\mathbf{a} - \mathbf{X}\beta\|^2 \\ &\quad + 2 \min_{\beta} \|\mathbf{b} - \mathbf{X}\beta\|^2 - \min_{\beta} \|\boldsymbol{\tau} - \mathbf{X}\beta\|^2, \end{aligned}$$

where $\boldsymbol{\tau} = \mathbf{a} - \mathbf{b}$ is the vector of all individual treatment effects.

Corollary 8.2. A tight upper bound on the limiting variance of the Horvitz–Thompson estimator under the Gram–Schmidt Walk design is

$$\text{VB} = \frac{2}{n} \min_{\beta} \|\mathbf{a} - \mathbf{X}\beta\|^2 + \frac{2}{n} \min_{\beta} \|\mathbf{b} - \mathbf{X}\beta\|^2 \geq V_{\text{GSW}}.$$

The fact that VB is an upper bound follows from $\min_{\beta} \|\tau - X\beta\|^2 \geq 0$. The fact that it is tight follows from that $\min_{\beta} \|\tau - X\beta\|^2 = 0$ when $\tau = \mathbf{0}$. If a constant is included among the covariates, the bound is tight whenever the treatment effects are constant, because then $\tau = \tau \mathbf{1}$ for some $\tau \in \mathbb{R}$. This mirrors the behavior of the Neyman variance bound (Neyman 1923). However, unlike the Neyman bound, the current bound is also tight whenever the covariates are perfectly predictive of the treatment effects.

To estimate VB, we first estimate $\beta_a = (X^T X)^{-1} X^T a$, and then plug it into an estimate of $\|a - X\beta_a\|^2 = \min_{\beta} \|a - X\beta\|^2$. Our assumptions ensure that β_a exists. Using Horvitz–Thompson-type estimators for both steps yields the estimator

$$\widehat{VB} = \frac{1}{n} \|\text{diag}(\mathbf{1} + z) (y - X\widehat{\beta}_a)\|^2 + \frac{1}{n} \|\text{diag}(\mathbf{1} - z) (y - X\widehat{\beta}_b)\|^2,$$

where $\widehat{\beta}_a = (X^T X)^{-1} X^T \text{diag}(\mathbf{1} + z)y$ and $\widehat{\beta}_b = (X^T X)^{-1} X^T \text{diag}(\mathbf{1} - z)y$. It is possible to use other estimators than Horvitz–Thompson-type estimators for these quantities, but we will not explore these alternative variance estimators in this article.

Because the variance bound is a quadratic form, the estimator is not unbiased, despite being based on the Horvitz–Thompson estimation principle. However, the estimator is consistent, as described in the following theorem.

Theorem 8.3. Suppose outcome and covariate regularity holds (Assumptions 7.1 and 7.2). Further suppose that the design parameter is bounded away from zero, $\phi = \Omega(1)$. Then, the variance bound estimator converges to the variance bound at the rate $\widehat{VB} - VB = \mathcal{O}_p(dn^{-1/2} \log(n))$.

When the number of covariates is bounded, $d = \mathcal{O}(1)$, the theorem states that the variance bound estimator is root- n consistent, up to a logarithmic factor. Under Assumption 7.4, stipulating that $d = o(n^{1/10-\epsilon})$, the convergence rate is somewhat slower at $n^{-(2/5+\epsilon)}$, again ignoring the logarithmic factor.

8.2. Confidence Intervals

Our confidence intervals are based on a normal approximation, motivated by Theorem 7.7. Let $\widehat{\sigma} = \widehat{VB}^{1/2}$ be the square root of the estimated variance bound, acting as a conservative estimator of the standard error of the treatment effect estimator. Furthermore, let $z_{\alpha} = \Phi^{-1}(1 - \alpha)$ be the tails of the standard normal distribution, where Φ^{-1} is its quantile function. A confidence interval at the $1 - \alpha$ confidence level is then given by endpoints $\widehat{\tau} \pm n^{-1/2} z_{\alpha/2} \widehat{\sigma}$.

Theorem 8.4. Suppose that Assumptions 7.1, 7.2, 7.3, and 7.4 hold. Further suppose that the design parameter approaches one at a sufficiently slow rate, so that $1 - \phi = o(1)$ and $1 - \phi = \omega(\xi^2/n)$. Then, the random interval centered at $\widehat{\tau}$ with radius $n^{-1/2} z_{\alpha/2} \widehat{\sigma}$ is an asymptotically valid $(1 - \alpha)$ -confidence interval:

$$\liminf_{n \rightarrow \infty} \Pr(-z_{\alpha/2} \widehat{\sigma} \leq n^{1/2} (\widehat{\tau} - \tau) \leq z_{\alpha/2} \widehat{\sigma}) \geq 1 - \alpha.$$

The confidence interval in Theorem 8.4 uses several asymptotic approximations. It is possible to modify the interval to

improve its finite-sample validity, at the cost of additional conservativeness. One important asymptotic approximation is that the interval is based on the limiting variance upper bound from Theorem 7.6, which showed that experimenters can escape the balance–robustness tradeoff asymptotically. However, it is not possible to escape the tradeoff in finite samples, so the limiting variance bound can be overly optimistic when the sample is small. In Section S8.5 in the supplement, we describe an alternative confidence interval with better finite-sample coverage, which is based on a more conservative variance estimator. This confidence interval is also valid when the design parameter ϕ does not approach one asymptotically. We discuss several other alternative confidence intervals in Section S8.5.

9. Comparison with Other Designs

9.1. Rerandomization

Rerandomization is a commonly used design approach to achieve covariate balance in experiments. The design is a uniform distribution over a set of assignment vectors that satisfy some acceptance criterion based on a measure of covariate balance, and it is often implemented by rejection sampling. Rerandomization implicitly navigates the balance–robustness tradeoff through the strictness of its acceptance criterion. In the version described by Lock Morgan and Rubin (2012), the acceptance criterion is based on the Mahalanobis distance between the means of covariates in the two treatment groups.

The properties of rerandomization are currently only well-understood in large samples, as described by Li, Ding, and Rubin (2018). Using Theorem 7.6, we can compare the limiting variance of the Gram–Schmidt Walk design with the limiting variance of rerandomization. In what follows, let $V_{CO} = n^{-1} \min_{\beta} \|\mu - \mathbf{1}\beta\|^2$ denote the limiting variance under complete randomization, let $V_{GSW} = n^{-1} \min_{\beta} \|\mu - X\beta\|^2$ denote the upper bound on the limiting variance of the Gram–Schmidt Walk design from Theorem 7.6, and let V_{RE} denote the limiting variance of rerandomization, as described by Li, Ding, and Rubin (2018).

Proposition 9.1. Suppose that Condition 1 in Li, Ding, and Rubin (2018) holds, that the second moment of the potential outcomes is asymptotically bounded, $n^{-1} \|\mu\|^2 = \mathcal{O}(1)$, and that a constant is included among the covariates, so that the first column of X is $\mathbf{1}$. Then, the limiting variance of the difference-in-means estimator under rerandomization, as described by Li, Ding, and Rubin (2018), is greater or equal to the limiting variance of the Horvitz–Thompson estimator under the Gram–Schmidt Walk design when ϕ satisfies the rate condition in Theorem 7.6: $V_{GSW} \leq V_{RE} \leq V_{CO}$. Equality holds, $V_{GSW} = V_{RE}$, only when $V_{RE} = V_{CO}$.

The proposition shows that the variance under the Gram–Schmidt Walk design always dominates the variance under rerandomization in large samples when a constant is included among the covariates. The only setting in which rerandomization and the Gram–Schmidt Walk design have the same limiting variance is when the covariates (excluding the constant) are completely uninformative of the potential outcomes. In that

case, the limiting variance under both designs is equal to the limiting variance under complete randomization.

Proposition 9.1 requires that a constant be included among the covariates to push the Gram–Schmidt Walk design toward treatment groups of equal sizes. Because the acceptance criterion of rerandomization is stated in terms of demeaned covariates and because it produces treatment groups of equal sizes by construction, its behavior is unchanged by the inclusion of the constant column. We discuss this further in Section S8.4.1 in the supplement. The proposition also requires that the second moment of the potential outcomes is bounded in addition to Condition 1 in Li, Ding, and Rubin (2018). This is because Li, Ding, and Rubin (2018) consider central moments in their analysis, while we consider raw moments.

The central insight underlying **Proposition 9.1** is that the limiting variance under rerandomization is a convex combination of the limiting variance under complete randomization and the limiting variance under Gram–Schmidt Walk design in **Theorem 7.6**. In particular, we show in the supplement that $V_{RE} = v_{K,a} V_{CO} + (1 - v_{K,a}) V_{GSW}$, where $v_{K,a} \in (0, 1)$ and $V_{CO} \geq V_{GSW}$. The coefficient $v_{K,a}$ is defined in Proposition 2 in Li, Ding, and Rubin (2018). It is the variance of a truncated random variable, which is shown to be the same as the ratio of the cumulative distribution functions of two Chi-squared random variables: $v_{K,a} = \Pr(\chi_{K+2}^2 \leq a) / \Pr(\chi_K^2 \leq a)$, where χ_K^2 denotes a Chi-squared random variable with degrees of freedom K . In the notation of Li, Ding, and Rubin (2018), K is the number of covariates, excluding the constant, and a is the balance acceptance threshold for the rerandomization procedure. As noted by Li, Ding, and Rubin (2018), it is an open question how to select a , but the authors suggest setting a so that $\Pr(\chi_K^2 \leq a) = 0.001$. Following this suggestion, we would have $v_{K,a} = 0.03$ when $K = 5$, meaning that rerandomization would be almost as asymptotically efficient as the Gram–Schmidt Walk design in that setting. However, we would have $v_{K,a} = 0.31$ when $K = 25$, which is a sizeable difference if the covariates are informative of the potential outcomes. That is, rerandomization yields less than 70% of the variance improvement over complete randomization compared to the Gram–Schmidt Walk design in this setting. Larger K makes this difference even more pronounced.

For rerandomization to achieve a limiting variance that is comparable to the Gram–Schmidt Walk design, experimenters must set the acceptance threshold a to be close to zero. Wang and Li (2022) provide a formal investigation along these lines. The authors show that when the acceptance criterion a approaches zero, meaning that experimenters reject an increasing share of drawn assignments as the sample grows, rerandomization achieves the same limiting variance as the Gram–Schmidt Walk design. However, setting the acceptance threshold close to zero will make the probability of accepting an assignment very small, often making the procedure infeasible to use in practice because computational resources are limited. For example, when $K = 25$, one needs to set $a = 0.27$ to achieve 99% of the improvement in asymptotic variance of the Gram–Schmidt Walk design, in the sense of $v_{K,a} = 0.01$. The probability of accepting an assignment is then less than 10^{-20} . The run time of the Gram–Schmidt Walk design is unaffected by the choice of design parameter ϕ .

9.2. Matched Pair Design

The matched pair design is another common experimental design to achieve covariate balance (Greevy et al. 2004). Units are here matched into pairs to minimize some objective function, which typically is the sum of Euclidean or Mahalanobis distances between the covariate vectors of paired units. After the pairs have been constructed, exactly one unit in each pair is assigned active treatment and the other unit control, independently between pairs.

The matched pair design achieves covariate balance by introducing dependence between paired units. This works well if paired units are nearly identical with respect to their covariates. The concern is that such nearly identical pairs are rare, even when matching on only a moderate number of covariates. Many, if not most, pairs will consist of units that are quite different from each other, and covariate balance will then not improve much despite considerable restrictions to randomization (and therefore robustness). For this reason, the matched pair design sacrifices a lot of robustness to achieve relatively little covariate balance, according to the operator norm measures. The following proposition formalizes this by considering randomly chosen covariate vectors to reflect typical problem instances. The argument implies there exist nonrandom vectors for which the same lower bound holds.

Proposition 9.2. Suppose n is an even integer and $\mathbf{x}_1, \dots, \mathbf{x}_n$ are drawn independently and uniformly from the d -dimensional unit ball with $d \geq 2$. For all matched pair designs,

$$\|\text{Cov}(\mathbf{z})\| = 2 \quad a.s., \quad \text{and} \quad \mathbb{E}[\xi^{-2} \|\text{Cov}(\mathbf{X}^T \mathbf{z})\|] \geq \frac{n^{1-2/d}}{8d},$$

where the covariances are taken with respect to the experimental design and the expectation is taken with respect to the random covariate vectors.

The proposition shows that there is a limit on the amount of covariate balance that can be achieved by the matched pair design. When there are $d \geq 3$ covariates, the lower bound on the imbalance grows with n . If $d = 10$, the best possible balance guarantee that the matched pair design can provide is $\gamma_x \geq \xi^2 n^{4/5} / 80$. This is better than the guarantee provided by the Bernoulli design, for which γ_x will be of order n , but it is not much better. The Gram–Schmidt Walk design provides the guarantee $\gamma_x = \xi^2 / (1 - \phi)$, regardless of n . The matched pair design has a fixed robustness guarantee of $\gamma_z = 2$, independent of the covariates and sample size. When $\phi = 1/2$, the Gram–Schmidt Walk design provides the guarantee $\gamma_z = 2$ and $\gamma_x = 2\xi^2$. So, the Gram–Schmidt Walk design with $\phi = 1/2$ provides better guarantees than the matched pair design on both robustness and covariate balance in large samples whenever $d \geq 3$.

Note that we are here comparing a lower bound for the matched pair design with an upper bound for the Gram–Schmidt Walk design. Even in situations where the lower bound for the matched pair design is lower than the upper bound for the Gram–Schmidt Walk design, the matched pair design will not necessarily provide better covariate balance.

The matched pair design differs from both the Gram–Schmidt Walk design and rerandomization by its targeting of all smooth functions of the covariates. This could be helpful if the

potential outcomes are explained well by nonlinear functions of the covariates. However, this comes at the cost of not providing much balance, neither on linear nor nonlinear functions. It is possible to emulate this behavior with the Gram–Schmidt Walk design by including an increasing number of transformations of the covariates as the sample grows. This has the benefit of being more targeted than trying to balance all aspects of the covariates all at once, but it also requires experimenters to make considered choices about which transformations to target. The same approach is not feasible with rerandomization, as acceptable assignments will be exceedingly rare when there are many things to balance, so it would take an insurmountable amount of time to find them using rejection sampling.

10. Additional Results and Extensions

We describe several extensions to the Gram–Schmidt Walk design and our analysis in the supplement. In Section S8.1, we relax the requirement of a symmetric design. With this relaxation, the experimenter provides a vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n) \in (0, 1)^n$ that specifies a desired assignment probability for each unit, and the design assigns the treatment accordingly. In Section S8.2, we derive finite-sample results for the Gram–Schmidt Walk design for matrix functions other than the operator norm, such as the trace norm and the Frobenius norm. In Section S8.3, we consider the balance–robustness tradeoff for other types of moment conditions than the bounded average magnitude condition, $PO(M)$.

In Section S8.4, we provide additional results on the sizes of the treatment groups under the Gram–Schmidt Walk design. We show that if $\phi < 1$ and at least a small constant is included among the covariates, the Gram–Schmidt Walk design will provide more balance on the group sizes than the Bernoulli design. However, the design will not ensure exact balance, in the sense of treatment groups that contain exactly $n/2$ units. We describe a modification of the design in Section S8.4 that achieves this, guaranteeing that treatment group sizes are exactly balanced. This modification breaks certain orthogonality properties of the updates of the algorithm, so our analysis does not apply to this modified design.

Section S10 of the supplement contains in-depth numerical illustrations of the behavior of the Gram–Schmidt Walk design and other commonly used designs. These simulations corroborate the theoretical results, showing that designs that balance the covariates well yield lower mean square error when the covariates are predictive, but are less robust. The Gram–Schmidt Walk design is shown to navigate the balance–robustness tradeoff well, providing more balance for a given level of robustness than other designs. The simulations also show that the confidence intervals cover the true average treatment effect at the nominal rate for moderate and large samples.

11. Concluding Remarks

Randomized experiments are useful because they provide robustness, but experimenters are often tempted to balance covariates with the aim of improving precision. The motivating idea of this article is that a compromise between balance and robustness is at the heart of the experimental design problem.

At one extreme, we can resolve this tradeoff cautiously by assigning treatments independently at random. This yields a design that is maximally robust. At the other extreme, we can make all assignments perfectly dependent. This yields a design that performs exceptionally well for some potential outcomes, but it will perform exceptionally poorly for other outcomes. Most experimenters would not prefer either of these extremes. Instead, they prefer intermediate designs that introduce weak dependencies between the assignments to achieve some balance at the cost of some robustness. The purpose of the Gram–Schmidt Walk design is to provide control over and efficiently navigate the tradeoff between covariate balance and robustness.

The question of which covariates should be balanced and how to trade off balance and robustness can only fully be answered by an experimenter’s preferences and substantive knowledge about the study at hand. In general terms, experimenters should prioritize balance over robustness, by setting the design parameter ϕ to a lower value, when they have access to high-quality covariates that are known to be predictive of the potential outcomes. Experimenters should also ensure that the set of covariates they balance is as linearly predictive as possible, by adding transformations and removing irrelevant covariates. We discuss practical considerations and heuristics related to the design in Section S1 in the supplement.

One of the chief short-comings of the Gram–Schmidt Walk design is that it solely focuses on linear functions. Experimenters can address this short-coming by balancing nonlinear transformation of the raw covariates, but this requires an active choice of which transformations to target. It is possible to extend the design to automatically balance nonlinear functions using kernel methods, but such an extension is beyond the scope of the current article. Another extension that is beyond the scope of the current article is an online version of Gram–Schmidt Walk design, where the experimenter must assign treatments to units in sequence without knowing the characteristics of future units.

Supplementary Materials

The supplementary materials contain proofs for all formal statements in the article, additional results and a simulation study. They also contain code to replicate the simulation study.

Acknowledgement

We thank Edo Airoldi, P. M. Aronow, Chen Chen, Nicholas Christakis, Peng Ding, Xavier D’Haultfœuille, Maximilian Kasy, Rad Niazadeh, David Pollard, Cyrus Samii, Jasjeet Sekhon and Johan Ugander for helpful comments and discussions. We thank Akshay Ramachandran for allowing us to include his proof of Lemma S3.5 in the supplement, which is shorter than our original proof.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported in part by NSF grant CCF-1562041, ONR Awards N00014-16-2374 and N00014-20-1-2335, a Simons Investigator Award to

Daniel Spielman. Christopher Harshaw was partially supported by NSF Graduate Research Fellowship (DGE1122492), NSF Foundations of Data Science Institute (FODSI) grant (DMS2023505), and funding from the Simons Institute for the Theory of Computing. Computing infrastructure was supplied by the Yale Center for Research Computing.

ORCID

Christopher Harshaw  <https://orcid.org/0000-0001-9350-8310>

Fredrik Sävje  <https://orcid.org/0000-0003-2544-8250>

References

- 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, 135–154. [2936]
- Aronow, P. M., and Samii, C. (2017), “Estimating Average Causal Effects under General Interference,” *Annals of Applied Statistics*, 11, 1912–1947. [2942]
- Bai, Y. (2023), “Why Randomize? Minimax Optimality under Permutation Invariance,” *Journal of Econometrics*, 232, 565–575. [2935]
- Banerjee, A. V., Chassang, S., Montero, S., and Snowberg, E. (2020), “A Theory of Experimenters: Robustness, Randomization, and Balance,” *American Economic Review*, 110, 1206–1230. [2935]
- Bansal, N., Dadush, D., Garg, S., and Lovett, S. (2019), “The Gram-Schmidt Walk: A Cure for the Banaszczyk Blues,” *Theory of Computing*, 15, 1–27. [2935,2938,2940]
- Basse, G. W., Ding, Y., and Toulis, P. (2022), “Minimax Designs for Causal Effects in Temporal Experiments with Treatment Habituation,” *Biometrika*, 110, 155–168. [2935]
- Beck, J., and Fiala, T. (1981), “Integer-Making Theorems,” *Discrete Applied Mathematics*, 3, 1–8. [2939]
- Bertsimas, D., Johnson, M., and Kallus, N. (2015), “The Power of Optimization over Randomization in Designing Experiments Involving Small Samples,” *Operations Research*, 63, 868–876. [2935]
- Bruhn, M., and McKenzie, D. (2009), “In Pursuit of Balance: Randomization in Practice in Development Field Experiments,” *American Economic Journal: Applied Economics*, 1, 200–232. [2935]
- Chatterjee, S., Dey, P. S., and Goswami, S. (2023), “Central Limit Theorem for Gram-Schmidt Random Walk Design,” arXiv:2305.12512. [2942]
- Cytrynbaum, M. (2021), “Designing Representative and Balanced Experiments by Local Randomization,” arXiv:2111.08157. [2935]
- Deaton, A., and Cartwright, N. (2018), “Understanding and Misunderstanding Randomized Controlled Trials,” *Social Science & Medicine*, 210, 2–21. [2935]
- Deville, J.-C., and Tillé, Y. (2004), “Efficient Balanced Sampling: The Cube Method,” *Biometrika*, 91, 893–912. [2939]
- Efron, B. (1971), “Forcing a Sequential Experiment to be Balanced,” *Biometrika*, 58, 403–417. [2934,2935,2936]
- Fisher, R. A. (1925), *Statistical Method for Research Workers*, Edinburgh: Oliver & Boyd. [2935]
- (1926), “The Arrangement of Field Experiments,” *Journal of the Ministry of Agriculture of Great Britain*, 33, 503–513. [2935]
- (1935), *The Design of Experiments*, London: Oliver & Boyd. [2935]
- Freedman, D. A. (1975), “On Tail Probabilities for Martingales,” *Annals of Probability*, 3, 100–118. [2940]
- Greevy, R., Lu, B., Silber, J. H., and Rosenbaum, P. (2004), “Optimal Multivariate Matching Before Randomization,” *Biostatistics*, 5, 263–275. [2935,2944]
- Higgins, M. J., Sävje, F., and Sekhon, J. S. (2016), “Improving Massive Experiments with Threshold Blocking,” *Proceedings of the National Academy of Sciences*, 113, 7369–7376. [2935]
- Imai, K., King, G., and Nall, C. (2009), “The Essential Role of Pair Matching in Cluster-Randomized Experiments,” *Statistical Science*, 24, 29–53. [2935]
- Johansson, P., Rubin, D. B., and Schultzberg, M. (2021), “On Optimal Rerandomization Designs,” *Journal of the Royal Statistical Society, Series B*, 83, 395–403. [2935]
- Kallus, N. (2018), “Optimal a Priori Balance in the Design of Controlled Experiments,” *Journal of the Royal Statistical Society, Series B*, 80, 85–112. [2935]
- Kapelner, A., Krieger, A. M., Sklar, M., and Azriel, D. (2022), “Optimal Rerandomization Designs via a Criterion that Provides Insurance Against Failed Experiments,” *Journal of Statistical Planning and Inference*, 219, 63–84. [2935]
- Kapelner, A., Krieger, A. M., Sklar, M., Shalit, U., and Azriel, D. (2021), “Harmonizing Optimized Designs with Classic Randomization in Experiments,” *The American Statistician*, 75, 195–206. [2935,2936]
- Kasy, M. (2016), “Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead,” *Political Analysis*, 24, 324–338. [2935]
- Krieger, A. M., Azriel, D., and Kapelner, A. (2019), “Nearly Random Designs with Greatly Improved Balance,” *Biometrika*, 106, 695–701. [2935]
- Leung, M. P. (2022), “Causal Inference Under Approximate Neighborhood Interference,” *Econometrica*, 90, 267–293. [2942]
- 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, 9157–9162. [2935,2943,2944]
- Lin, W. (2013), “Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman’s Critique,” *Annals of Applied Statistics*, 7, 295–318. [2942]
- Lock Morgan, K., and Rubin, D. B. (2012), “Rerandomization to Improve Covariate Balance in Experiments,” *Annals of Statistics*, 40, 1263–1282. [2935,2943]
- McLeish, D. L. (1974), “Dependent Central Limit Theorems and Invariance Principles,” *The Annals of Probability*, 2, 620–628. [2942]
- Neyman, J. (1923), “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9,” *Statistical Science*, 5, 465–472. Reprinted in 1990. [2943]
- Nordin, M., and Schultzberg, M. (2022), “Properties of Restricted Randomization with Implications for Experimental Design,” *Journal of Causal Inference*, 10, 227–245. [2935]
- Spencer, J. (1985), “Six Standard Deviations Suffice,” *Transactions of the American Mathematical Society*, 289, 679–679. [2935]
- Student. (1938), “Comparison between Balanced and Random Arrangements of Field Plots,” *Biometrika*, 29, 363–378. [2935]
- Wang, Y., and Li, X. (2022), “Rerandomization with Diminishing Covariate Imbalance and Diverging Number of Covariates,” *Annals of Statistics*, 50, 3439–3465. [2944]
- Wu, C.-F. (1981), “On the Robustness and Efficiency of Some Randomized Designs,” *Annals of Statistics*, 9, 1168–1177. [2935]