CS 270 Final Project: Gram–Schmidt Walk

Christina Jin, Jerry Lai

May 7, 2021

Contents

1 Introduction 2
1.1 Balance-Robustness Trade-off . . . 2
1.2 Randomized Experiment Setup and Notations . . . 2

2 Quantifying robustness and covariate balance 3
2.1 Spectral Decomposition of the Mean Square Error . . . 3
2.2 Baseline: Minimax Optimal Design . . . 3
2.3 Integrating Covariate Balance . . . 3

3 Gram-Schmidt Walk Design 4
3.1 Design Introduction . . . 4
3.1.1 Augmented Covariates . . . 4
3.1.2 The Gram-Schmidt Walk Design . . . 4
3.2 Design Properties and Performance compared to baseline . . . 5
3.2.1 Bound on Robustness . . . 5
3.2.2 Bound on Covariate Balance . . . 5
3.2.3 Bound on Mean Square Error . . . 6

4 Gram-Schmidt Walk Algorithm 7
4.1 Algorithm Description . . . 7
4.1.1 Explanation of Variables and Notation . . . 7
4.2 Algorithm Analysis . . . 8
4.2.1 Covariate Balance . . . 8
4.2.2 Unbiasedness . . . 9

5 Conclusion 9

6 Bibliography 10

7 Appendix 11
7.1 Proof of Theorem 3.2.1 . . . 11
7.2 Proof of Theorem 3.2.2 . . . 11
7.3 Proof of Theorem 3.2.4 . . . 12

1 Introduction

1.1 Balance-Robustness Trade-off

The algorithm covered by this paper, the Gram-Schmidt Walk, is used in the field of experimental design for causal inference, specifically in randomized control trials. In these trials, samples are assigned to different treatment groups. This placement of samples can be done with different methodologies, with varying degrees of randomness. These methodologies need to consider two key metrics that are used to judge the assignments: **Covariate Balance** and **Robustness**.

Covariate Balance is important because it ensures that the only difference between treatment groups is the treatment itself. One way to achieve good covariate balance is to deliberately group samples in such a way that balances observable covariates. However, doing this may lead to bias and imbalances among the unobservable covariates that are worse than what would be achieved with random assignments.

Robustness is how well can ensure that there are no systematic errors, or bias, in the assignment of treatment. In other words, the most robust methods ensure that the group assignments are balanced *in expectation*, under any circumstances. An example of a highly robust design method is fully random selection. However, fully random selection may not always be balanced for particular assignments, and it can't prevent unsystematic errors.

The **Gram-Schmidt Walk** design attempts to address the problem of the balance-robustness trade-off by giving researchers control over how much emphasis to put on the two metrics and maximizing covariate balance subject to a prespecified level of robustness.

1.2 Randomized Experiment Setup and Notations

In the causal inference setting for this report, we consider an experiment with n units, a binary treatment variable $\boldsymbol{z} \in \mathbb{R}^n$, and an outcome variable $\boldsymbol{y} \in \mathbb{R}^n$. We denote the assignment to unit i as $z_i \in \{\pm 1\}$. The observed outcome for unit i is:

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

The quantity of interest in this report is the *average treatment effect (ATE)*:

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

Note that τ is unobservable because we could only either a_i or b_i for unit i (but not both). To estimate τ, we consider the Horvitz–Thompson estimator, which is known to be unbiased and consistent for many designs (Narain, 1951; Horvitz & Thompson, 1952). For designs where $\mathbb{P}(z_i = 1) = \frac{1}{2}$, this estimator could be written as:

$$\widehat{\tau} = \frac{1}{n} \sum_{i:z_i=1} \frac{y_i}{0.5} - \frac{1}{n} \sum_{i:z_i=-1} \frac{y_i}{0.5} = \frac{2}{n} \langle \boldsymbol{z}, \boldsymbol{y} \rangle$$

2 Quantifying robustness and covariate balance

2.1 Spectral Decomposition of the Mean Square Error

Let $\boldsymbol{\mu} = (\boldsymbol{a} + \boldsymbol{b})/2$ be the average of the potential outcome vectors. Let $\boldsymbol{\eta}_i$ be the i^{th} normalized eigenvector of $\mathrm{Cov}(\boldsymbol{z})$ with eigenvalue λ_i.

Lemma 2.1.1 *For any experimental design with $\mathbb{P}(z_i = 1) = 1/2\ \forall i \in [n]$, the mean square error of the Horvitz–Thompson estimator is:*

$$\mathbb{E}\left[(\widehat{\tau} - \tau)^2\right] = \frac{4}{n^2}\boldsymbol{\mu}^\top \mathrm{Cov}(\boldsymbol{z})\boldsymbol{\mu} \tag{1}$$

$$= \frac{4M}{n}\sum_{i=1}^{n} w_i^2 \lambda_i \tag{2}$$

where in (2) $M = \frac{1}{n}\sum_{i=1}^{n}\mu_i^2$ is the second moment of $\boldsymbol{\mu}$, and $w_i = \langle \boldsymbol{\mu}, \boldsymbol{\eta}_i\rangle^2/\|\boldsymbol{\mu}\|^2$ is the alignment of $\boldsymbol{\mu}$ with $\boldsymbol{\eta}_i$.

We see from Equation (2) that the MSE is proportional to a convex combination of the eignevalues. To minimize the MSE, a design needs to "target" the eigenvectors that are close to the directions of potential outcome vector by making their corresponding eigenvalues small. Therefore, a design is a bet on the direction of the potential outcome vector.

2.2 Baseline: Minimax Optimal Design

The worst-case MSE is $4\lambda_{max}M/n$, when $\boldsymbol{\mu}$ is parallel with the eigenvector corresponding to the largest eigenvalue λ_{max}. As a result, the most robust design, or the safest bet, is one that focuses equally on all possible directions of the potential outcome vector, i.e.: $\lambda_i = 1\ \forall i \in [n]$. Such a design achieves the smallest worst-case MSE ($=\frac{4M}{n}$), and is therefore known as the "minimax optimal design". This is achieved by any design where the assignments are pair-wise independent, i.e.: $\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{I}$.

2.3 Integrating Covariate Balance

Let $\boldsymbol{X}$ be the $n \times d$ covariate matrix whose rows are covariate vectors for each unit $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. If the covariate vectors are predictive of the potential outcomes, we may better better minimize MSE by balancing covariates.

Formally, let $\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}\|$. Then we could define $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\boldsymbol{\beta}$ as the predicted potential outcomes and $\boldsymbol{\varepsilon} = \boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}$ as the errors of the prediction. Plugging this into equation (2), we have:

$$\frac{n^2}{4}\mathbb{E}\left[(\widehat{\tau} - \tau)^2\right] = \underbrace{\widehat{\boldsymbol{\mu}}^\top \mathrm{Cov}(\boldsymbol{z})\widehat{\boldsymbol{\mu}}}_{\text{term 1}} + \underbrace{\boldsymbol{\varepsilon}^\top \mathrm{Cov}(\boldsymbol{z})\boldsymbol{\varepsilon}}_{\text{term 2}} + 2\widehat{\boldsymbol{\mu}}^\top \mathrm{Cov}(\boldsymbol{z})\boldsymbol{\varepsilon}$$

where term 1 could be re-written as $\boldsymbol{\beta}^\top \mathrm{Cov}(\boldsymbol{X}^\top \boldsymbol{z})\boldsymbol{\beta}$. Note that $\mathrm{Cov}(\boldsymbol{X}^\top \boldsymbol{z})$ is capturing the covariate imbalances.

To see the balance-robustness tradeoff, note that $\widehat{\boldsymbol{\mu}} \perp \boldsymbol{\varepsilon}$, so we couldn't minimize both terms simultaneously. However, if our prior knowledge suggests that the covariates are predictive, then we could shift our aim towards minimizing the spectral norm of $\mathrm{Cov}(\boldsymbol{X}^T\boldsymbol{z})$, i.e.: aligning $\mathrm{Cov}(\boldsymbol{z})$ disproportionally with the $\mathbf{Column}(\boldsymbol{X})$. On the other hand, if the covariates are not predicative, we should focus on maximizing robustness by making $\mathrm{Cov}(\boldsymbol{z})$ small in all directions, i.e.: not aligning $\mathrm{Cov}(\boldsymbol{z})$ in any particular direction.

3 Gram-Schmidt Walk Design

3.1 Design Introduction

The goal of the design is to give experimenters control over the balance-robustness trade-off mentioned above by introducing a parameter $\phi \in [0, 1]$. For a desired level of robustness, the design tries to maximize balance. This means that we want to make the covariance balance matrix $\mathrm{Cov}(\boldsymbol{X}^T\boldsymbol{z})$ as small as possible, and thus simultaneously balance all linear functions of the covariates. ($\boldsymbol{z} \in \mathbb{R}^n$ is the assignment vector of treatment, and $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is the raw covariate matrix.)

3.1.1 Augmented Covariates

One input of the design is the matrix of *augmented covariates* $\boldsymbol{B} \in \mathbb{R}^{(n+d)\times n}$, constructed as follows:

$$\boldsymbol{B} = \begin{bmatrix} \sqrt{\phi}\boldsymbol{I} \\ \xi^{-1}\sqrt{1-\phi}\boldsymbol{X}^\top \end{bmatrix}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\xi = \max_{i\in[n]} ||x_i||$ is the max row norm of the covariate matrix $\boldsymbol{X}$. The factor ξ ensures that the two constituents are on comparable scales. We can see why ϕ controls the trade-off here. If $\phi = 0$, the augmented covariates resemble the raw covariates $\boldsymbol{X}^T$, and the design places more emphasis on covariate balance. In particular, it will minimize the spectral norm of $\mathrm{Cov}(\boldsymbol{X}^T\boldsymbol{z})$. If $\phi = 1$, the augmented covariates completely ignore the raw covariates and the algorithm output will be the minimax optimal mentioned in section 2.2.

3.1.2 The Gram-Schmidt Walk Design

The Gram-Schmidt Walk design involves calling the Gram-Schmidt Walk algorithm with input vector $\boldsymbol{z}_1$ and matrix $\boldsymbol{B}$. The goal of the algorithm is to generate $\boldsymbol{z} \in \{\pm 1\}^n$ such that $\boldsymbol{Bz} \approx 0$. We will describe the algorithm in more details in Section 4, so we will just treat it as a black box here. The output vector $\boldsymbol{z}$ will be the treatment units assignments.

The augmented covariate vector of unit i, denoted as $\boldsymbol{b}_i \in \mathbb{R}^{(n+d)}$, is the ith column of $\boldsymbol{B}$:

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

where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ is the ith n-dimensional basis vector. Since $\boldsymbol{Bz}$ could be expressed as follows, it is also known as the difference between the within-group sums:

$$\boldsymbol{Bz} = \sum_{i=1}^{n} z_i\boldsymbol{b}_i = \sum_{i:z_i=1} \boldsymbol{b}_i - \sum_{i:z_i=-1} \boldsymbol{b}_i.$$

3.2 Design Properties and Performance compared to baseline

3.2.1 Bound on Robustness

Theorem 3.2.1 *Under the Gram-Schmidt Walk design, the worst-case mean squared error is upper bounded by the ratio between the minimax optimum and the design parameter ϕ. In other words, for all potential average outcome vectors $\boldsymbol{\mu} = (\boldsymbol{a} + \boldsymbol{b})/2$, all covariate matrices $\boldsymbol{X}$, and all parameter values $\phi \in (0, 1]$,*

$$\mathbb{E}[(\hat{\tau} - \tau)^2] \leq \frac{4M}{\phi n} = \frac{1}{\phi} MSE_{minimax} \quad where \quad M = \frac{1}{n}\sum_{i=1}^{n} \mu_i^2$$

Proof: See Appendix 7.1. □

This theorem shows that we could attain a desired robustness guarantee by picking an appropriate ϕ, in the worst case, where there is no association between the covariates and the outcomes. We also see no dependency on the covariate dimension for this worst-case bound, meaning that the design has a certain level of performance guarantee in high-dimensional regimes.

3.2.2 Bound on Covariate Balance

Theorem 3.2.2 *Under the Gram-Schmidt Walk design, the imbalance of any linear function $\boldsymbol{v} = \boldsymbol{X\theta}$ of the covariates is bounded by:*

$$\mathbb{E}\left[(\boldsymbol{v}^\top \boldsymbol{z})^2\right] \leq \frac{\xi^2 \|\boldsymbol{\theta}\|^2}{1 - \phi}$$

Proof: See Appendix 7.2. □

Note that the worst-case bound of Theorem 3.2.2 decreases monotonically with ϕ, indicating less imbalance as $\phi \to 0$.

To compare the above theorem with the baseline (from Section 2.2), we have the following corollary:

Corollary 3.2.3 *Suppose $\xi = \mathcal{O}(\sqrt{d \log(n)})$ and $\|\boldsymbol{\theta}\|^2 / \|\boldsymbol{X\theta}\|^2 = \mathcal{O}(1/n)$, the covariate imbalance relative to the baseline (independent assignment) is:*

$$\frac{\mathbb{E}\left[\langle \boldsymbol{v}, \boldsymbol{z}_{\text{gsw}} \rangle^2\right]}{\mathbb{E}\left[\langle \boldsymbol{v}, \boldsymbol{z}_{ind} \rangle^2\right]} = \mathcal{O}\left(\frac{d \log(n)}{(1 - \phi)n}\right)$$

Proof: $\mathbb{E}\left[\langle \boldsymbol{v}, \boldsymbol{z}_{\text{ind}} \rangle^2\right] = \mathbb{E}\left[\left(\boldsymbol{\theta}^\top \boldsymbol{X}^\top \boldsymbol{z}_{\text{ind}}\right)^2\right] = \|\boldsymbol{X\theta}\|^2$. The corollary directly follows by plugging the values specified in the condition. □

The first condition on xi makes sure that the covariates don't have extreme outliers asymptotically. The second condition means that we don't look at $\boldsymbol{\theta}$ that is trivially balanced by all designs, e.g.: $\boldsymbol{\theta} = 0$.

This corollary tells us that as long as $\phi \neq 1$, meaning that we ensure the design to at least partially focus on achieving balance, the relative imbalance between the GSW design and the baseline design is on the order $\mathcal{O}\left(\frac{d}{n}\right)$ (disregarding the log factor).

3.2.3 Bound on Mean Square Error

To look at the joint effect of robustness and covariate balance, we could examine the bound on the mean square error, which is our ultimate goal of the design.

Theorem 3.2.4 *The mean squared error under the GSW design is at most the minimum of the loss function of an implicit ridge regression of the average of the potential outcome vectors* $\boldsymbol{\mu}$ *on the covariates:*

$$\mathbb{E}\left[(\widehat{\tau}-\tau)^2\right] \leq \frac{4L}{n} \qquad \textit{where} \qquad L = \min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\frac{1}{\phi n}\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\xi^2}{(1-\phi)n}\|\boldsymbol{\beta}\|^2\right].$$

Proof: See Appendix 7.3. □

This theorem suggests that the MSE depends on the predictiveness of the covariates for the potential outcomes, captured by L, the optimal loss of a ridge regression with a regularization parameter of $\xi^2\phi/(1-\phi)$. The loss is scaled by $1/\phi$, which suggests one shortcoming of this bound: it could be quite conservative for small ϕ.

The first term of the loss captures how well $\boldsymbol{\beta}$ predicts the potential outcomes using the covariates. The second term captures the magnitude of $\boldsymbol{\beta}$, scaled by ξ^2 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. As $\phi \to 0$, the optimal β converges to the best linear predictor, and thus the design prioritizes covariate balance.

Combining this with Theorem 3.2.1, we see that the worst-case occurs when $L = M/\phi = \|\mu\|^2/\phi n$. From the definition of L, we see that this occurs when $\boldsymbol{\beta} = \mathbf{0}$. This confirms our intuition from above that the worst case is when the covariates are completely unpredictive.

4 Gram-Schmidt Walk Algorithm

4.1 Algorithm Description

Algorithm 1 Gram Schmidt Walk

```
Input: $\boldsymbol{z}_1 = \mathbf{0}$, and $\boldsymbol{B} \in \mathbb{R}^{(n+d)\times n}$, matrix of augemented covariates
Output: $\boldsymbol{z} \in \{-1, 1\}^n$, final vector of assignments
Set iteration index $t \leftarrow 1$ and alive set $A_1 \leftarrow [1 \text{ to } n]$.
while $\mathcal{A}_t \neq \emptyset$ do
    $p_t \leftarrow$ the largest index in $\mathcal{A}_t$
    $\boldsymbol{u}_t \leftarrow \underset{\boldsymbol{u}\in\mathbb{R}^n}{\arg\min} \|\boldsymbol{B}\boldsymbol{u}\| \ \ s.t. \ \ (\boldsymbol{u}[p_t] = 1) \wedge (\boldsymbol{u}[i] = 0 \ \forall i \notin \mathcal{A}_t)$
    $\Delta \leftarrow \{\delta \in \mathbb{R} : \boldsymbol{z}_t + \delta \boldsymbol{u}_t \in [-1, 1]^n\}$          ▷ We know $0 \in \Delta$
    $\delta_t^+ \leftarrow |\max \Delta|$ and $\delta_t^- \leftarrow |\min \Delta|$          ▷ $\delta_t^+ = (\max \Delta)$ and $\delta_t^- = -(\min \Delta)$
    Set step size $\delta_t$ at random according to
        $\delta_t \leftarrow \begin{cases} \delta_t^+ \text{ with probability } \delta_t^- / \left(\delta_t^+ + \delta_t^-\right) \\ -\delta_t^- \text{ with probability } \delta_t^+ / \left(\delta_t^+ + \delta_t^-\right) \end{cases}$
    Update the fractional assignment $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t + \delta_t \boldsymbol{u}_t$
    Update the alive set $\mathcal{A}_{t+1} \leftarrow \{i \mid \boldsymbol{z}_t[i] \neq \pm 1\}$
    $t \leftarrow t + 1$
end while
Return $\boldsymbol{z}$
```

4.1.1 Explanation of Variables and Notation

- $\boldsymbol{B} \in \mathbb{R}^{(n+d)\times n}$: the matrix of augmented covariates
- $\boldsymbol{z_t} \in [-1, 1]^n$: the vector of assignments, updated with each iteration of the algorithm. The initial vector, $\boldsymbol{z}_1$, is an input to the algorithm and is initialized to $\boldsymbol{z}_1 = \mathbf{0}$. At the end of the algorithm, $\boldsymbol{z} \in \{-1, 1\}^n$, represents the group assignments for each of the n samples.
- $\mathcal{A}_t$: the *alive set.* At any time step t, $\mathcal{A}_t$ is the set of indices of $\boldsymbol{z}_t$ for which the entry is not +1 or -1. $\mathcal{A}_1$ contains all n indices.

  At each iteration, only the indices in $\mathcal{A}_t$ are updated for $\boldsymbol{z}_t$. The other entries in $\boldsymbol{z}_t$, which are already +1 or -1, keep their value until the end.
- $\boldsymbol{u}_t \in \mathbb{R}^n$: The step direction for updating $\boldsymbol{z}_t$ at time t
- p_t, the pivot, is one of the indices in $\mathcal{A}_t$. In this version of the algorithm, it is set to be the last index in $\mathcal{A}_t$; other formulations can have different choices for the pivot. When computing the step direction, the pivot's entry in the step direction vector is constrained to $\boldsymbol{u}[p_t] = 1$.
- δ_t: The step size for updating $\boldsymbol{z}_t$ at time t

4.2 Algorithm Analysis

In the introduction, we stated that this algorithm maximizes covariate balance subject to a specified level of robustness (or unbiasedness). The following subsections detail how the algorithm achieves this, first with covariate balance then with unbiasedness.

4.2.1 Covariate Balance

This algorithm handles the balancing problem as a relaxation. Whereas the original problem is defined as group assignments where every value is integral, i.e. every entry in the vector is $\{\pm 1\}$, this algorithm handles it in a fractional form where each entry is in $[-1, 1]$.

The algorithm's purpose is to push the assignments to become integral, while maintaining an acceptable level of balance.

At the beginning of the algorithm, the balance between covariates, expressed by $\boldsymbol{Bz}$, is perfect, because

$$\boldsymbol{Bz}_1 = \boldsymbol{B0} = \boldsymbol{0}$$

However, this state is obviously not acceptable as a solution because the entries of $\boldsymbol{z}_1$ are not integral. As $\boldsymbol{z}$ gets pushed closer and closer to being integral, the balance becomes harder to maintain, as $\boldsymbol{Bz}$ becomes harder to keep close to $\boldsymbol{0}$ as $\boldsymbol{z}$ moves further away from $\boldsymbol{0}$.

At each time step, the update is determined by a direction $\boldsymbol{u}$ and step size δ:

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

The step direction is determined to minimize the change in balance of $\boldsymbol{z}$ by minimizing the imbalance of the update direction itself by minimizing $\|\boldsymbol{Bu}_t\|$. The rationale for this can be explained by modifying the above espression to:

$$\boldsymbol{Bz}_{t+1} = \boldsymbol{Bz}_t + \delta_t \boldsymbol{Bu}_t$$

The minimization of $\|\boldsymbol{Bu}_t\|$ is subject to two constraints.

1. $\boldsymbol{u}_t[i] = 0 \;\forall i \notin \mathcal{A}_t$

   This constraint simply ensures that only the elements in the alive set $\mathcal{A}_t$ are updated at each time step. This is because the elements that are not in $\mathcal{A}_t$ are already integral, and should no longer be updated.

2. $\boldsymbol{u}_t[p_t] = 1$, where p_t is the pivot, the last element in the alive set. This is done for two reasons:

   (a) To avoid the trivial solution to minimize $\|\boldsymbol{Bu}_t\|$, $\boldsymbol{u}_t = \boldsymbol{0}$

   (b) To avoid compounding imbalances in the updates. [insert explanation on pivot phases]

The choice of step size must lie in Δ, the set of all possible step sizes where $\boldsymbol{z}_t + \delta \boldsymbol{u}_t \in [-1, 1]^n$. δ_t is chosen randomly between δ_t^+ and $-\delta_t^-$, which are the positive and negative extrema of Δ respectively. The purpose of choosing an extremum is to ensure that at each iteration, at least one element of $\boldsymbol{z}$ in $\mathcal{A}_t$ is pushed to become integral and thus leave $\mathcal{A}_t$. Since the indices that have left the alive set will never go back in again, this puts an upper bound on the number of iterations in this algorithm.

4.2.2 Unbiasedness

We will show the unbiasdness of the GSW algorithm with Theorem 4.2.1 and the following Corollary.

Theorem 4.2.1 *Under the Gram-Schmidt Walk algorithm,* $\mathbb{E}[\boldsymbol{z}_t] = \boldsymbol{z}_1\ \forall t$.

Proof: Consider the conditional expection of the update rule $\boldsymbol{z}_{t+1} \leftarrow \boldsymbol{z}_t + \delta_t \boldsymbol{u}_t$:

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

From Law of Iterated Expectations:

$$\mathbb{E}[\delta_t] = \mathbb{E}[\mathbb{E}[\delta_t \mid \delta_t^+, \delta_t^-]]$$

Since δ_t is conditionally independent of all other variables given (δ_t^+, δ_t^-):

$$\mathbb{E}[\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t] = \boldsymbol{z}_t + \mathbb{E}[\mathbb{E}[\delta_t \mid \delta_t^+, \delta_t^-]\boldsymbol{u}_t \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t] \tag{3}$$

In the algorithm, we randomly select δ_t with the following probabilities:

$$\delta_t \leftarrow \begin{cases} \delta_t^+ \text{ with probability } \delta_t^- / \left(\delta_t^+ + \delta_t^-\right) \\ -\delta_t^- \text{ with probability } \delta_t^+ / \left(\delta_t^+ + \delta_t^-\right) \end{cases}$$

We can use this to calculate $\mathrm{E}[\delta_t]$.

$$\mathbb{E}[\delta_t \mid \delta_t^+, \delta_t^-] = \delta_t^+ \left(\frac{\delta_t^-}{\delta_t^+ + \delta_t^-}\right) - \delta_t^- \left(\frac{\delta_t^+}{\delta_t^+ + \delta_t^-}\right) = 0 \tag{4}$$

Substituting Equation 4 into Equation 3, we get:

$$\mathbb{E}[\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_1, \ldots, \boldsymbol{z}_t] = \boldsymbol{z}_t + \boldsymbol{0}$$

$$\forall t,\ \mathbb{E}[\boldsymbol{z}_t] = \boldsymbol{z}_1$$

□

Theorem 4.2.1 shows that, if we set $\boldsymbol{z}_1 \leftarrow \boldsymbol{0}$ as we do in the design, then the expectation of the final assignments is $\boldsymbol{0}$.

It follows that $\forall i \in [n],\ \mathbb{P}[\boldsymbol{z}[i] = 1] = \mathbb{P}[\boldsymbol{z}[i] = -1] = \frac{1}{2}$.

Thus proving that the Gram-Schmidt algorithm does not introduce any bias as it pushes each assignment towards being integral, and that assignments for each element are uniformly random.

5 Conclusion

The aim of this report is to explain the balance-robustness tradeoff problem in the field of experimental design, and provide an algorithm for experimenters to control this tradeoff through its parameter. As shown in the analysis of the properties of the design, it provides guarantee on both robustness and balance.

There are also some shortcomings of the GSW design that motivates some open questions in the field. One open question is that since the GSW design proposed here solely focuses on linear functions of the covariates, could we find an algorithm that also balances non-linear functions of the covariates? Another open question is to look at whether it's possible to upper bound the MSE in terms of the covariance matrix of the imbalances of the augmented covariates, $\mathrm{Cov}(\boldsymbol{B}^{\top}\boldsymbol{z})$, directly, instead of using the upper bound L, as shown in section 3.2.4, so that we could provide a tighter bound regardless of ϕ.

6 Bibliography

- Christopher Harshaw, Fredrik Sävje, Daniel A. Spielman, and Peng Zhang (2021). Balancing covariates in randomized experiments with the Gram–Schmidt Walk design
  `https://arxiv.org/pdf/1911.03071.pdf`
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer, second edition.
- Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663–685.
  `https://www.jstor.org/stable/2280784`
- Narain, R. D. (1951). On sampling without replacement with varying probabilities. Journal of the Indian Society of Agricultural Statistics, 3(2), 169-175.
- Nikhil Bansal, Daniel Dadush, Shashwat Garg, Shachar Lovett (2017). The Gram-Schmidt Walk: A cure for the Banaszczyk Blues
  `https://arxiv.org/pdf/1708.01079.pdf`

7 Appendix

We need the following theorem in our proof for some of the main results in this survey.

Theorem 7.0.1 *Under the Gram-Schmidt Walk design,*

$$\mathrm{Cov}(\boldsymbol{Bz}) \preceq \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T.$$

7.1 Proof of Theorem 3.2.1

Proof: As mentioned in section 2.2, the worst-case MSE of any design is $4\lambda_z M/n$, where λ_z is the largest eigenvalue of $\mathrm{Cov}(\boldsymbol{z})$. We thus need to show that $\lambda_z \leq 1/\phi$.

If we restrict matrix $\mathrm{Cov}(\boldsymbol{Bz})$ to its upper left $n \times n$ block, $\phi\mathrm{Cov}(\boldsymbol{z})$, and apply Theorem 7.0.1, it follows that

$$\phi\mathrm{Cov}(\boldsymbol{z}) \preceq \phi\boldsymbol{Q} \quad \text{where} \quad \boldsymbol{Q} = \left(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}.$$

This implies that $\lambda_z \leq \lambda_{max}(\boldsymbol{Q})$. Since $\boldsymbol{Q}$ is positive definite and thus invertible for all ϕ, $\lambda_{max}(\boldsymbol{Q}) = 1/\lambda_{min}(\boldsymbol{Q^{-1}})$. Observe that:

$$\begin{aligned}\lambda_{min}(\boldsymbol{Q^{-1}}) &= \lambda_{min}(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\top) \\ &= \phi + (1-\phi)\xi^{-2}\lambda_{min}(\boldsymbol{X}\boldsymbol{X}^\top) \\ &\geq \phi & (\boldsymbol{X}\boldsymbol{X}^\top \text{ is P.S.D.})\end{aligned}$$

Therefore, $\lambda_z \leq \lambda_{max}(\boldsymbol{Q}) = 1/\lambda_{min}(\boldsymbol{Q^{-1}}) \leq 1/\phi$. □

7.2 Proof of Theorem 3.2.2

Proof: If we restrict matrix $\mathrm{Cov}(\boldsymbol{Bz})$ to its lower right $d \times d$ block, $\xi^{-2}(1-\phi)\mathrm{Cov}(\boldsymbol{X}^\top\boldsymbol{z})$, and apply Theorem 7.0.1, it follows that

$$\xi^{-2}(1-\phi)\mathrm{Cov}(\boldsymbol{X}^\top\boldsymbol{z}) \preceq \xi^{-2}(1-\phi)\boldsymbol{X}^\top\left(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X}$$

Diving both sides by $\xi^{-2}(1-\phi)$, we have

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{X}^\top\boldsymbol{z}) &\preceq \boldsymbol{X}^\top\left(\phi\boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{X} \\ &= \frac{\xi^2}{1-\phi}\underbrace{\boldsymbol{X}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top + \frac{\xi^2\phi}{1-\phi}\boldsymbol{I}\right)^{-1}\boldsymbol{X}}_{=\boldsymbol{H}}\end{aligned}$$

Thus, for any linear function $\boldsymbol{\theta} \in \mathbb{R}^d$ of the covariates, we have

$$\begin{aligned}\boldsymbol{\theta}^\top\mathrm{Cov}(\boldsymbol{X}^\top\boldsymbol{z})\boldsymbol{\theta} &\leq \frac{\xi^2}{1-\phi}\boldsymbol{\theta}^\top\boldsymbol{H}\boldsymbol{\theta} \\ \mathbb{E}\left[\left(\boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{z}\right)^2\right] &\leq \frac{\xi^2}{1-\phi}\boldsymbol{\theta}^\top\boldsymbol{H}\boldsymbol{\theta}\end{aligned}$$

Since $\boldsymbol{H}$ could be interpreted as the hat matrix for a ridge regression with $\frac{\xi^2}{1-\phi}$ as the regularization parameter, all of its eigenvalues are at most 1. This implies that $\boldsymbol{\theta}^\top\boldsymbol{H}\boldsymbol{\theta} \leq \|\theta\|^2$, $\forall\boldsymbol{\theta} \in \mathbb{R}^d$.
□

7.3 Proof of Theorem 3.2.4

Proof: In the proof of Theorem 3.2.1, we showed that $\mathrm{Cov}(\boldsymbol{z}) \preceq \boldsymbol{Q} = \left(\phi \boldsymbol{I} + (1-\phi)\xi^{-2}\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}$. Using this and Lemma 2.1.1, we have

$$\mathbb{E}\left[(\widehat{\tau}-\tau)^2\right] = \frac{4}{n^2}\boldsymbol{\mu}^\top \mathrm{Cov}(\boldsymbol{z})\boldsymbol{\mu} \le \frac{4}{n^2}\boldsymbol{\mu}^\top \boldsymbol{Q}\boldsymbol{\mu}.$$

All we need to show now is that $nL = \min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\frac{1}{\phi}\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\xi^2}{1-\phi}\|\boldsymbol{\beta}\|^2\right] = \boldsymbol{\mu}^\top \boldsymbol{Q}\boldsymbol{\mu}$. First, note that the minimization problem has a closed-form solution $\boldsymbol{\beta^*}$ as follows:

$$\begin{aligned}
\boldsymbol{\beta}^* &= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\frac{1}{\phi}\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\xi^2}{1-\phi}\|\boldsymbol{\beta}\|^2\right] \\
&= \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d}\left[\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{\xi^2\phi}{1-\phi}\|\boldsymbol{\beta}\|^2\right] \\
&= \left(\underbrace{\boldsymbol{X}^\top\boldsymbol{X} + \frac{\xi^2\phi}{1-\phi}\boldsymbol{I}}_{=\boldsymbol{R}}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\mu} \\
&= \boldsymbol{R}^{-1}\boldsymbol{X}^\top\boldsymbol{\mu}
\end{aligned}$$

Now we can plug $\boldsymbol{\beta}^*$ into the objective function to get:

$$nL = \frac{1}{\phi}\|\boldsymbol{\mu}-\boldsymbol{X}\boldsymbol{\beta^*}\|^2 + \frac{\xi^2}{1-\phi}\|\boldsymbol{\beta^*}\|^2 = \frac{1}{\phi}\boldsymbol{\mu}^\top\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\top\right)\boldsymbol{\mu}$$

We are leaving out the algebraic details here, but it's straightforward and simplifies well!

To complete the proof, we apply the Woodbury Identity: $(\boldsymbol{I}+\boldsymbol{U}\boldsymbol{C}\boldsymbol{V})^{-1} = \boldsymbol{I} - \boldsymbol{U}(\boldsymbol{C}^{-1} + \boldsymbol{V}\boldsymbol{U})^{-1}\boldsymbol{V}$ with $\boldsymbol{U}=\boldsymbol{X}, \boldsymbol{V}=\boldsymbol{X}^\top$, and $\boldsymbol{C} = \frac{1-\phi}{\xi^2\phi}\boldsymbol{I}$. As a result, we get

$$\begin{aligned}
\frac{1}{\phi}\left(\boldsymbol{I}-\boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{X}^\top\right) &= \frac{1}{\phi}\left(\boldsymbol{I}+\boldsymbol{X}\frac{\xi^{-2}(1-\phi)}{\phi}\boldsymbol{I}\boldsymbol{X}^\top\right)^{-1} \\
&= \frac{1}{\phi}\left(\boldsymbol{I}+\frac{\xi^{-2}(1-\phi)}{\phi}\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \\
&= \left(\phi\boldsymbol{I}+\xi^{-2}(1-\phi)\boldsymbol{X}\boldsymbol{X}^\top\right)^{-1} \\
&= \boldsymbol{Q}
\end{aligned}$$

Thus we have $nL = \boldsymbol{\mu}^\top\boldsymbol{Q}\boldsymbol{\mu}$. □