On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference

Nathan Kallus

Abstract
I study the minimax-optimal design for a two-arm controlled experiment where conditional mean outcomes vary in a given set and the objective is effect-estimation precision. When this set is permutation symmetric, the optimal design is shown to be complete randomization. Notably, even when the set has structure (i.e., is not permutation symmetric), being minimax-optimal for precision still requires randomization beyond a single partition of units, that is, beyond randomizing the identity of treatment. A single partition is not optimal even when conditional means are linear. Since this only targets precision, it may nonetheless not ensure sufficient uniformity for design-based (i.e., randomization) inference. I therefore propose the inference-constrained mixed-strategy optimal design as the minimax-optimal for precision among designs subject to sufficient-uniformity constraints.

KEYWORDS
causal inference, controlled experimentation, covariate balance, design of experiments, optimization, permutation test

1 INTRODUCTION

A common criticism of completely randomized designs for controlled experiments is that they may result in assignments that appear ‘imbalanced’. Student (1938) said it ‘would be pedantic to continue with an arrangement of [field] plots known beforehand to likely lead to a misleading conclusion’. Both the judgement of ‘imbalance’ and the supposed foreknowledge of misleading conclusions, however, must depend on some knowledge of how post-treatment outcomes depend on pre-treatment variables.
Recently, Kallus (2018) developed a systematic framework of how such knowledge translates to optimal design. The framework is phrased as a zero-sum game against Nature, where the experimenter seeks to eliminate noise and Nature interferes adversarially but is constrained by structure assumed on the conditional expectation function of post-treatment outcomes given pre-treatment variables. In this paper, I revisit this framework to highlight the minimax optimality of randomization, even if we focus on just a variance objective and not only randomization inference. Nonetheless, even though it has randomization, it may not have enough to ensure enough conditional power for randomization tests. I therefore propose a new constrained version that does.

I also use this framing to discuss Johansson et al. (2021) who recently compared rerandomization (Morgan & Rubin, 2012) and the pure-strategy optimal design (PSOD) of Kallus (2018), which is a computational heuristic offered for the minimax-optimal mixed-strategy optimal design (MSOD). I thank and congratulate Johansson et al. (2021) for a thought-provoking paper, and in particular for highlighting the importance of enabling randomization inference in addition to minimizing estimation variance. I use the opportunity to set straight a few facts: randomization beyond treatment blinding is in fact minimax-variance optimal; designing for optimal precision subject to enabling randomization inference does not require uniform randomization over a restricted set; and finally Theorem 1 (‘no free lunch’) and Example 1 of Kallus (2018) are correct (and appear to have been misinterpreted by Johansson et al., 2021) in showing that, in the worst-case, randomizing between two symmetric assignments, and in particular optimizing the Mahalanobis distance between experimental group means, can increase variance by a factor of $n - 1$ relative to complete randomization.

2 | THE FRAMEWORK OF KALLUS (2018) REDUX AND REFINED

We briefly review the framework of Kallus (2018), focusing on two treatment arms. Our sample consists of $n$ units with (observed) pretreatment variables $X_i \in \mathcal{X}$ and (unobserved) potential outcomes $Y_i(1), Y_i(-1) \in \mathbb{R}$, for $i = 1, \ldots, n$. Each unit $(X_i, Y_i(1), Y_i(-1))$ is assumed independent of others. Define $\mu_i = \mathbb{E}[Y_i(1) + Y_i(-1)|X_i]$ and $\epsilon_i = Y_i(1) + Y_i(-1) - \mu_i$. We are interested in estimating and making inferences on the sample average treatment effect (SATE): $\tau = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(-1))$.

Towards that end, we choose treatment assignments $W = (W_1, \ldots, W_n) \in \{-1,1\}^n$ and observe $Y_i^{\text{obs}} = Y_i(W_i)$. (We use boldface throughout to refer to $n$-tuples.) We restrict to $n$ even and $W^T \mathbf{1} = 0$, where $\mathbf{1} = (1, \ldots, 1)$. As outcomes are not observed before treatment, $W$ must be independent of $(Y(-1), Y(1))$ given $X$. We focus on the estimator $\hat{\tau} = \frac{1}{n} W^T Y^{\text{obs}}$.

A design is a distribution $\sigma$ over $W$, specifying how we assign treatments conditional on $X$; we treat $\sigma$ as a random variable measurable with respect to $X$. We require that, under $\sigma$, $W$ and $-W$ are equiprobable. We refer to this property as blinding the identity of treatment. Specifically, $\sigma$ satisfies

$$\sigma \in \mathcal{D} = \{ \mathbb{R}_+^n : \sum_{W \in \mathcal{W}} \sigma(W) = 1, \sigma(W) = \sigma(-W) \forall W \in \mathcal{W} \}$$

Given $X$ and a design $\sigma$, $(Y(-1), Y(1), W, Y^{\text{obs}})$ have a joint distribution (conditional on $X$).

We can show that, since $\mathbb{E}[\hat{\tau}|X, Y(-1), Y(1)] = \tau$ by blinding, $\hat{\tau} - \tau = \frac{1}{n} W^T Y(-1) + Y(1)$ by algebra, and $\mathbb{E}[\epsilon_i|X] = \mathbb{E}[\epsilon_i\epsilon_j|X] = 0$ for $i \neq j$ by independence, we have (Kallus, 2018, Theorem 7)

$$\text{var}[\hat{\tau}] = \frac{1}{n^2} \mathbb{E}[V(\sigma, \mu)] + \frac{1}{n^2} \sum_{i=1}^{n} \text{var}[\epsilon_i] + \text{var}[\tau],$$

where $V(\sigma, \mu_0) = \sum_{W \in \mathcal{W}} \sigma(W)(W^T \mu_0)^2$. 

(1)
We do not know \( \mu \) so we consider a minimax framework. Given \( X \) and some set \( \mathcal{M} \subseteq \mathbb{R}^n \) of potential values for \( \mu \), we define \( V(\sigma, \mathcal{M}) = \sup_{\mu \in \mathcal{M}} V(\sigma, \mu) \). A minimax-optimal design is one minimizing \( V(\sigma, \mathcal{M}) \) (for every \( X \)). This is called the MSOD in Kallus (2018) to emphasize that it is a mixed strategy in this zero-sum game, that is it randomizes over unit partitions. A heuristic approximation, the PSOD is defined as the uniform randomization over the set \( \arg\min_{W \in \mathcal{W}} V(WQ^T, \mathcal{M}) \), which is shown to recover blocking and nonbipartite-pair-matched designs for some specific choices of \( \mathcal{M} \).

3 | THE OPTIMALITY OF COMPLETE RANDOMIZATION

A natural next question is, what is the minimax-optimal design? That, of course, depends on \( \mathcal{M} \). If we have no particular knowledge about \( \mu \), we should not constrain \( \mathcal{M} \) in any informative fashion. However, \( V(\sigma, \mu_0) \) is linear in \( \mu_0 \) so we must restrict \( \mathcal{M} \) somehow else \( V(\sigma, \mathcal{M}) = \infty \) (equivalently, we measure \( V(\sigma, \mu_0) \) relative to a magnitude of \( \mu_0 \)). An uninformative restriction must be permutation symmetric: invariant to permutations of the coordinates of \( \mathbb{R}^n \). An exercise in convexity shows that, whenever \( \mathcal{M} \) is permutation symmetric, complete randomization over \( \mathcal{W} \) is a minimax-optimal design. This is the ‘no free lunch’ theorem of Kallus (2018, Theorem 1): one cannot improve upon complete randomization unless one is willing to assume structure that is a deviation from symmetry.

One important permutation symmetric example of \( \mathcal{M} \) is \( \mathcal{M}_{CR} = \{ \mu_0 : V(\sigma_{CR}, \mu_0) \leq C \} \), where \( \sigma_{CR} \) denotes the complete randomization (CR) design. Then \( V(\sigma, \mathcal{M}_{CR}) \) exactly measures the variance under \( \sigma \) relative to that under complete randomization.

4 | THE SUBOPTIMALITY OF A SINGLE ASSIGNMENT

Consider a blinded design using a single partition of units: \( \sigma_S(W_0) = \sigma_S(-W_0) = \frac{1}{2} \) for some \( W_0 \in \mathcal{W} \). Since \( Q(\sigma_S) = W_0W_0^T \), \( Q(\sigma_{CR}) = \frac{n}{n-1}(I - \frac{W_0W_0^T}{n}) \), we have \( V(\sigma_S, \mathcal{M}_{CR}) = C(n-1) \). This says that, for any \( W_0 \), there always exists \( \mu_0 \in \mathbb{R}^n \) with \( V(\sigma_{CR}, \mu_0) = C, V(\sigma_S, \mu_0) = C(n-1) \).

Taking \( C = \Theta(1/n) \), we see a single partition can have non-vanishing worst-case variance, while CR always ensures \( 1/n \) variance.

Example 1 of Kallus (2018) provides an example construction where such a single partition minimizes any scaled Euclidean distance between group means in \( X \) (such as Mahalanobis distance): \( b \in \mathbb{N}, \ n = 2^b, \ X_i = \sum_{t=0}^{b-\max\left\{2\log_2(i+1)\right\}} (-1)^t \cdot 2^{b-1+t} + i \mod 2^b \), \( \mu_i = (-1)^i \). Then \( \arg\min_{W \in \mathcal{W}} |W^TX| = \{ W_0,-W_0 \} \) where \( W_0 = (-1)^i \), yielding \( V(\sigma_{CR}, \mu_0) = 4/(n-1), V(\sigma_S, \mu_0) = 4 \). The example is also appealing as it recovers the worst-case behaviour of nonbipartite pair matching and blocking. Johansson et al. (2021) appear to incorrectly cite this example as ‘\( X = (2^0, \ldots, 2^{n/2-1}, -2^0, \ldots, -2^{n/2-1}) \)’, for which \( \{ \pm W_0 \} \) are not the only optimizers of \( |W^TX| \). But Example 1 of Kallus (2018) actually says that the above ‘complicated construction essentially yields \( X \approx \text{round}(X) = (2^0, \ldots, 2^{n/2-1}, -2^0, \ldots, -2^{n/2-1}) \) with just enough perturbation so that the assignment \( \pm W_0 \) uniquely minimizes Mahalanobis distance between group means’. The claim of Johansson et al. (2021) that ‘The mistake of Kallus (2018) stems from the incorrect assumption that the allocation \( \pm W_0 \) uniquely minimizes the Mahalanobis distance for all \( n \)’ is false: they consider a different set of covariates than Kallus (2018). More generally, per the computation in the preceding paragraph, whenever some \( \pm W_0 \)
uniquely optimize the Mahalanobis distance up to signs (which happens generically when covariates are randomly drawn), there will always exist some mean-outcome vector \( \mu \) such that the design \( \sigma_{\text{Maha-opt}} \) randomizing over all optimizers of the Mahalanobis distance between group means will have \( V(\sigma_{\text{Maha-opt}}, \mu) / V(\sigma_{\text{CR}}, \mu) = n - 1 \). Example 1 of Kallus (2018) is just one (correct) explicit example.

### 5 | THE OPTIMALITY OF RESTRICTED RANDOMIZATION: THE MIXED-STRATEGY OPTIMAL DESIGN

When is something different from CR minimax-optimal? That still depends on \( \mathcal{M} \). An example is \( \mathcal{M} = \mathcal{M}_K = \{ Ku: u^T Ku \leq C \} \) for a positive semidefinite \( K \), for which \( V(\sigma, \mathcal{M}_K) = C \lambda_{\text{max}}(K^{1/2} Q(\sigma) K^{1/2}) \) where \( \lambda_{\text{max}}(\cdot) \) denotes the largest eigenvalue. The optimal designed is therefore determined by the spectrum of \( K \), except for on \( I \), which is always in \text{null}(Q(\sigma)). \)

If the remaining spectrum (of multiplicity \( n - 1 \)) is concentrated at a single value then \( \mathcal{M}_K \) is permutation symmetric and CR is optimal. At the other extreme, if \( K = uu^T \) is of rank one, then using any single partition \( \pm W_0 \in \arg\min_{W \in \mathcal{W}} |W^T u| \) becomes the optimal design. In between these extremes, when \( K \) has a dispersed spectrum, something between perfectly partitioning a single vector and complete randomization is optimal: the (randomized) MSOD proposed in Kallus (2018), which solves \( \min_{\sigma \in \mathcal{H}} \lambda_{\text{max}}(K^{1/2} Q(\sigma) K^{1/2}) \). Kallus (2018) gives tractable inner and outer semidefinite-programming approximations to this generally hard problem.

We may, for example, take \( K_{ij} = \mathcal{H}(X_i, X_j) \) for a kernel \( \mathcal{H} \), corresponding to assuming \( \mu_i = f(X_i) \forall i \) for \( f \) varying in the unit ball of the reproducing kernel Hilbert space (RKHS) given by \( \mathcal{H} \). This offers the researcher a flexible modelling framework, clearly connecting assumed structure on \( f \) to an optimal design. As one simple example when \( X_i \in \mathbb{R}^d \), we can take linear functions \( f \in \{ x \mapsto \beta^T x : \beta^T \Omega^{-1} \beta \leq 1 \} \), which corresponds to taking \( \mathcal{H}(x,x') = x^T \Omega x' \). The single-design partition has \( V(\sigma_{\text{opt}}, \mathcal{M}_K) = D_\Omega(W_0) := W_0^T X \Omega X^T W_0 \), so the optimal single-partition design, that is the PSOD, minimizes the distance in pretreatment group means—Mahalanobis distance if \( \Omega \) is the inverse sample covariance. However, if \( d \geq 2 \) then the resulting \( K \) is generically of rank higher than one and a single partition will not be minimax-optimal. This means that, unlike the characterization of Johansson et al. (2021), even in the simple linear setting that recovers Mahalanobis mean-matching, a single partition is not optimal—we need more randomization.

### 6 | RERANDOMIZATION: SUBOPTIMAL BUT BUILT FOR RANDOMIZATION INFERENCE

Morgan and Rubin (2012) propose uniform randomization over \( \{ W \in \mathcal{W} : D_\Omega(W) \leq a \} \), operationalized by repeatedly uniformly sampling \( W \in \mathcal{W} \) until \( D_\Omega(W) \leq a \), termed rerandomization. In our minimax framework the rerandomization design is, however, not minimax-optimal, even for \( \mathcal{M} = \mathcal{M}_X \Omega X^T \). The only exception is the case \( d = 1 \), where a single partition is optimal and we can use \( a = \min_{W \in \mathcal{W}} D_\Omega(W) \) to recover it (i.e., the PSOD). In practice, usually \( d \geq 2 \), in which case the minimax-optimal design (i.e., the MSOD) requires more randomization than a single partition, but the rerandomization design is generally not minimax-optimal for any value of \( a \).

Nonetheless, as Johansson et al. (2021) highlight, rerandomization crucially facilitates randomization inference while improving precision (even if suboptimally): we need only ensure
The MSOD already has randomization, even beyond a single partition, and we can always apply a randomization test to any design. A concern, however, is the power. One might hope that high precision implies high power, but if we randomize over a single partition we can easily have high precision while for two-sided statistics a randomization test will always return $p = 1$ and we never reject the null. We must therefore ensure that each assignment only occurs with probability at most $\alpha$ for any design. In contrast, for any design $\sigma$ that has $\sigma(W_0) = \sigma(-W_0) > \frac{\alpha}{2}$ for some $W_0$, a randomization test with a two-sided statistic will never reject the null at significance $\alpha$ whenever $W_0$ is the chosen assignment. Unfortunately, if we set $N_a = \frac{\alpha}{2}$, rerandomization sampling is virtually interminable ($N_a/|\mathcal{W}| = 2^{-\Theta(n)}$) and we also still do not obtain a minimax-optimal design. And, if we fix the acceptance probability $N_a/|\mathcal{W}| \approx F_{\chi^2}(a)$ as constant in $n$ (e.g., 0.1), we will randomize over very many partitions with very large imbalance $D_{\Omega}(W)$.

7  |  Optimizing for Randomization Inference

The MSOD already has randomization, even beyond a single partition, and we can always apply a randomization test to any design. A concern, however, is the power. One might hope that high precision implies high power, but if we randomize over a single partition we can easily have high precision while for two-sided statistics a randomization test will always return $p = 1$ and we never reject the null. We must therefore ensure that each assignment only occurs with probability at most $\alpha/2$ (focusing on two-sided statistics). The MSOD, despite being randomized, may or may not have this property for any one given $\alpha$.

I therefore propose the inference-constrained MSOD, which, given $K$ and $\alpha$, solves:

$$
\min_{\sigma \in \mathcal{D}}: \sigma(W) \leq \frac{\alpha}{2} \forall W \in \mathcal{W} \lambda_{\max}(\sum_{W \in \mathcal{W}} \sigma(W) K^{1/2} W W^T K^{1/2}).
$$

(2)

The constraint $\sigma(W) \leq \frac{\alpha}{2}$ ensures that each assignment has probability at most $\alpha/2$ so that, if chosen, the randomization test can potentially return $p \leq \alpha$ if the statistic is indeed extreme. Equation (2) can be a difficult optimization problem so I propose the following approximation based on Kallus (2018, Algorithm 3). Set $\mathcal{W}_1 = \mathcal{W} \cap \{ W: W_1 = 1 \}$; for $t = 1, \ldots, T$, solve $W_t \in \arg\min_{W \in \mathcal{W}_t} W^T K W$ and set $\mathcal{W}_{t+1} = \mathcal{W}_t \cap \{ W: W_t^T W \leq n - 4 \}$. Then $W_1, -W_1, \ldots, W_T, -W_T$ are the top $2T$ solutions to $\min_{W \in \mathcal{W}} W^T K W$ (the top two, if unique, gives the PSOD). Each optimization problem is a binary optimization problem with a convex-quadratic objective and linear constraints and can be solved with off-the-shelf solvers such as Gurobi. Then let $U \in \mathbb{R}^{n \times T}$ have the columns $W_t$ and solve

$$
\min_{\lambda \in \mathbb{R}, u \in \mathbb{R}^n_+} : u \leq \frac{\alpha}{2}, u^T 1 = 1, \lambda - K^{1/2} U \text{diag}(u) U^T K^{1/2} \text{is positive semidefinite}, \lambda,
$$

(3)

and set $\sigma(W_t) = \sigma(-W_t) = u_t/2$. Equation (3) is a tractable semidefinite program. Notice we need at least $T \geq 1/\alpha$ for Equation (3) to be feasible. If $1/\alpha$ is integral, setting $T = 1/\alpha$ forces Equation (3) to choose the design that uniformly randomizes over the top $T$ solutions to $\min_{W \in \mathcal{W}} W^T K W$. This latter alternative approach was considered in Kallus (2018, Example 4) but was found empirically less powerful than a bootstrap test. Focusing solely on randomization tests, Equation (2) is exactly the minimax-optimal design for optimizing variance subject to the constraint of no single assignment occurring more often than $\alpha/2$.

For larger $T$, Equation (3) provides a good approximation of Equation (2).

8  |  Concluding Remarks

For variance, we showed it is generally minimax optimal to randomize beyond the identity of treatment. Neither a single partition nor rerandomization is optimal. I thank Johansson et al. (2021) for
highlighting the importance of considering randomization inference as an additional objective. A principled approach to multiple objectives is to optimize one while constraining the other. Rather than rerandomization over $2/\alpha$ assignments, this suggests seeking the minimax-optimal design over those with sufficiently uniform randomization to enable $\alpha$-significance design-based testing after randomization.

REFERENCES
Johansson, P., Rubin, D.B. & Schultzberg, M. (2021) On optimal rerandomization designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2), 395–403.
Kallus, N. (2018) Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1), 85–112.
Morgan, K.L. & Rubin, D.B. (2012) Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2), 1263–1282.
Student. (1938) Comparison between balanced and random arrangements of field plots. *Biometrika*, 29(3/4), 363–378.

How to cite this article: Kallus N. On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference. *J R Stat Soc Series B*. 2021;83:404–409. [https://doi.org/10.1111/rssb.12412](https://doi.org/10.1111/rssb.12412)