Submitted to the Annals of Statistics
arXiv: arXiv:1312.0531

OPTIMAL A PRIORI BALANCE IN THE DESIGN OF CONTROLLED EXPERIMENTS

By Nathan Kallus

Massachusetts Institute of Technology

We develop a unified theory of designs for controlled experiments that balance baseline covariates a priori (before treatment and before randomization) using the framework of minimax variance. We establish a "no free lunch" theorem that indicates that, without structural information on the dependence of potential outcomes on baseline covariates, complete randomization is optimal. Restricting the structure of dependence, either parametrically or non-parametrically, leads directly to imbalance metrics and optimal designs. Certain choices of this structure recover known imbalance metrics and designs previously developed ad hoc, including randomized block designs, pairwise-matched designs, and re-randomization. New choices of structure based on reproducing kernel Hilbert spaces lead to new methods, both parametric and non-parametric. *

1. Introduction. Achieving balance between experimental groups is a corner stone of causal inference, otherwise any observed difference may be attributed to a difference other than the treatment alone. In clinical trials, and more generally controlled experiments, where the experimenter controls the administration of treatment, complete randomization of subjects has been the golden standard for achieving this balance on average.

The expediency of complete randomization, however, has been controversial since the founding of statistical inference in controlled experiments. William Gosset, "Student" of Student's T-test, said of assigning field plots to agricultural interventions that it "would be pedantic to continue with an arrangement of [field] plots known beforehand to be likely to lead to a misleading conclusion," such as arrangements in which one experimental group is on average higher on what he calls the "fertility slope" than the other experimental group [1]. Of course, as the opposite is just as likely under complete randomization, this is not an issue of estimation bias in its modern definition, but of estimation variance. Gosset's sentiment is echoed in

MSC 2010 subject classifications: Primary 62K05; secondary 90C99
Keywords and phrases: Optimal experimental design, controlled experiments, covariate balance
*This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374.

the common statistical maxim "block what you can, randomize what you cannot" attributed to George Box and in the words of such individuals as James Heckman ("Randomization is a metaphor and not an ideal or 'gold standard"' [2]) and Donald Rubin ("For gold standard answers, complete randomization may not be good enough" [3]). In one interpretation, these can be seen as calls for the experimenter to ensure experimental groups are balanced at the onset of the experiment, before applying treatments and before randomization.

There is a variety of designs for controlled experiments that attempt to achieve better balance in terms of measurements made prior to treatment, known as baseline covariates, under the understanding that a predictive relationship possibly holds between baseline covariates and the outcomes of treatment. We term this sort of approach *a priori balancing* as it is done before applying treatments and before randomization (the term a priori is chosen to contrast with post hoc methods such as post stratification, which may be applied after randomization and after treatment [4]). The most notable a priori balancing designs are randomized block designs [5], pairwise matching [6], and re-randomization [7].[1]

Each of these implicitly defines imbalance between experimental groups differently. Blocking attempts to achieve exact matching (when possible): a binary measure of imbalance that is zero only if the experimental groups are identical in their discrete or coarsened baseline covariates. Pairwise matching treats imbalance as the sum of pairwise distances, given some pairwise distance metric such as Mahalanobis. There are both globally optimal and greedy heuristic methods that address this imbalance measure [12]. In [7], the authors define imbalance as the group-wise Mahalanobis distance and propose re-randomization as a heuristic method for reducing it non-optimally.

It is unclear when each of these different characterizations of imbalance is appropriate and when is deviating from complete randomization justified. The connection between an imbalance metric such as the sum of pairwise distances before treatment and estimation variance after treatment is also unclear. We here argue that, without structural information on the dependence of outcomes on baseline covariates, complete randomization is minimax optimal. Furthermore, when structural knowledge is expressed as membership of conditional expectations in a normed vector space of functions, an alternative minimax-optimal rule arises for the a priori balancing of experimental

[1]There are also sequential methods to address the case where allocation must be decided before all subjects are admitted [8, 9, 10]. These are beyond the present scope of this paper. Response-adaptive designs that use outcome data to inform future assignments (see [11]) lie between a priori and post hoc and are also beyond our scope.

groups. We show how certain choices of this structure reconstruct each of the aforementioned methods or associated imbalance metrics. We study other choices of structure using reproducing kernel Hilbert spaces (RKHS), which give rise to new methods, both parametric and non-parametric.

We study in generality the characteristics of any such method that arises from our framework, including its estimation variance and consistency, intimately connecting a priori balance to post-treatment estimation. Whenever a parametric model of dependence is known to hold, we show that, relative to complete randomization, the variance due to the optimal design converges linearly ($2^{-\Omega(n)}$ for n subjects) to the best theoretically possible – a generalization of the observation on linear convergence made in [13]. We provide algorithms for finding the optimal designs using mixed integer optimization (MIO) and semi-definite optimization (SDO) and hypothesis tests that are appropriate for these designs. We make connections to Bayesian experimental design and shed light on the usefulness of a priori balance in designing experiments plagued by non-compliance.

1.1. *Structure of the paper.* In Section 2, we consider the effect of structure and the lack thereof. In particular, we set up the problem, argue that complete randomization is optimal in the absence of structural information (Section 2.1), define structural information and the resulting imbalance metrics and optimal designs (Section 2.2), show how this recovers existing imbalance metrics and designs (Section 2.3), study the designs that arise from RKHS structure (Section 2.4), and consider a Bayesian interpretation (Section 2.4.1). We end Section 2 with simulation studies of fictitious data (Example 2.2) and of clinical data (Example 2.3). In Section 3, we characterize the variance (Section 3.1), consistency (Section 3.2), and rate of convergence (Section 3.3) of estimators arising from a priori balancing designs. In Section 4, we provide algorithms for finding the optimal designs. In Section 5, we provide hypothesis tests for making inferences on treatment effects. We offer some concluding remarks in Section 6.

All proofs are given in the supplement. In the supplement, we also consider the benefit of a priori balancing to experiments plagued by non-compliance (Section 7.1) and generalizations of structural information (Section 7.2).

2. The effect of structural information and lack thereof. We begin by describing the set up. Let m denote the number of treatments to be investigated (including controls). We index the subjects by $i = 1, \ldots, n$ and assume $n = mp$ is divisible by m. We assume the subjects are independently randomly sampled but we will consider estimating both sample and population effects. We denote assigning subject i to a treatment k by $W_i = k$.

We let $w_{ik} = \mathbb{I}[W_i = k]$ and $W = (W_1, \ldots, W_n)$. When $m = 2$, we will use $u_i = w_{i1} - w_{i2}$. As is common for controlled trials, we assume non-interference (see e.g. [14, 15] and p. 19 of [16]). I.e., a subject assigned to a certain treatment exhibits the same outcome regardless of others' assignments. Under this assumption we are able to define the potential post-treatment outcome Y_{ik} of subject i were it to be subjected to the treatment k. We let Y denote the matrix of all potential outcomes. We assume throughout Y_{ik} has second moments. Let X_i, taking values in some $\mathcal{X}$, be the baseline covariates of subject i that are recorded before treatment and let $X = (X_1, \ldots, X_n)$.

We denote by $\mathrm{TE}_{kk'i} = Y_{ik} - Y_{ik'}$ the unobservable causal treatment effect for subject i. There are two unobservable quantities that will be of interest to estimate. One is the *sample average (causal) treatment effect* (SATE):

$$\mathrm{SATE}_{kk'} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{TE}_{kk'i} = \frac{1}{n}\sum_{i=1}^{n} Y_{ik} - \frac{1}{n}\sum_{i=1}^{n} Y_{ik'}.$$

Another is the *population average (causal) treatment effect* (PATE):

$$\mathrm{PATE}_{kk'} = \mathbb{E}[\mathrm{TE}_{kk'1}] = \mathbb{E}[\mathrm{SATE}_{kk'}].$$

By construction, SATE is an unbiased and strongly consistent estimate of PATE. Our estimator will always be the simple mean differences estimator

$$\hat{\tau}_{kk'} = \frac{\sum_{i:W_i=k} Y_{ik}}{\sum_{i:W_i=k} 1} - \frac{\sum_{i:W_i=k'} Y_{ik'}}{\sum_{i:W_i=k'} 1}.$$

We drop subscripts when $m = 2$ and set $k = 1$, $k' = 2$.

Throughout we will consider only designs that

(2.1) do not depend on future information, that is, W is independent of Y, conditional on X;

(2.2) blind (randomize) the identity of treatments, that is, $\mathbb{P}(W = (k_1, \ldots, k_n) \,|X) = \mathbb{P}(W = (\pi(k_1), \ldots, \pi(k_n)) \,|X)$ for any permutation π of $1, \ldots, m$; and

(2.3) split the sample evenly, that is, surely $\sum_{i:W_i=k} 1 = p \quad \forall k$.

We interpret conditions (2.1)-(2.3) as the definition of *a priori balance* as they require that all balancing be done before applying treatments (condition (2.1)) and before randomization (conditions (2.2)-(2.3)). Condition (2.1) is a reflection of the temporal logic of first assigning, then experimenting. Condition (2.2) says that balancing is done before randomization and it ensures that the estimators $\hat{\tau}_{kk'}$ resulting from the design are always unbiased, both conditionally on X, Y (i.e., in estimating SATE) and marginally

(i.e., in estimating PATE; more detail given in Theorem 3.1). Condition (2.3) is a way to achieve (2.2) in non-completely-randomized designs. If W is an even assignment then randomly permuting treatment indices will blind their identity. Else, given one fixed uneven assignment, a treatment can be identified by the size of its experimental group.

We denote by $\mathcal{W} \subset \{1, \dots, m\}^n$ the space of feasible assignments satisfying (2.3) and by $\Delta \subset [0,1]^{|\mathcal{W}|}$ the space of feasible designs (distributions over assignments) satisfying (2.1)-(2.3). For $m = 2$ we also write $\mathcal{W} \cong \mathcal{U} = \{u \in \{-1, +1\}^n : \sum_i u_i = 0\}$ and $\mathcal{P} = \text{covex-hull}(\mathcal{U})$.

2.1. *No free lunch.* We will now argue that without structural information on the relationship between X_i and Y_{ik}, complete randomization is minimax optimal. For the rest of this subsection we will restrict to $m = 2$.

Among estimators that are unbiased, the standard way of comparing efficiency is variance. By the law of total variance and by the conditional unbiasedness of any estimator resulting from a design satisfying (2.1)-(2.3),

$$\text{Var}\left(\hat{\tau}\right) = \mathbb{E}\left[\text{Var}\left(\hat{\tau}|X, Y\right)\right] + \text{Var}\left(\text{SATE}\right).$$

The variance of SATE is independent of our choice of a priori balancing design. This choice can only affect the first term. Therefore, an efficient design will seek to minimize $\text{Var}\left(\hat{\tau}|X, Y\right)$ path-by-path, i.e. for the given subjects at hand. Whatever the design does to minimize this term will not affect the second term as long as the design adheres to the above conditions.

Denote by $\hat{\tau}^{\text{CR}}$ the estimator arising from complete randomization, which randomizes uniformly over equal partitions independently of X. Then,

$$\text{Var}\left(\hat{\tau}^{\text{CR}}|X, Y\right) = \frac{4}{n(n-1)} \left|\left|\overline{Y}\right|\right|_2^2$$
$$\text{where } \hat{Y}_i = \frac{Y_{i1} + Y_{i2}}{2},\ \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i,\ \text{and } \overline{Y}_i = \hat{Y}_i - \hat{\mu}.$$

Using this as a benchmark, we compare efficiency based on the normalized unitless ratio $\text{Var}\left(\hat{\tau}|X, Y\right) / \text{Var}\left(\hat{\tau}^{\text{CR}}|X, Y\right)$.

However, we do not know Y, only X (condition (2.1)), and we assume no structural information on their relationship. Therefore, we consider an adversarial Nature that chooses Y so to increase our variance. The following shows that in this situation, complete randomization is optimal.

THEOREM 2.1. *Fix $X \in \mathcal{X}^n$. Let $||\cdot||$ be any permutationally invariant seminorm on $\mathbb{R}^n$. Then, among designs satisfying (2.1)-(2.3) (i.e., among*

all $\sigma \in \Delta$), complete randomization minimizes either of

$$\max_{Y\in\mathbb{R}^{n\times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\|Y_1\|^2+\|Y_2\|^2} = \max_{Y\in\mathbb{R}^{n\times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\|\hat{Y}\|^2} = \max_{Y\in\mathbb{R}^{n\times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\|\overline{Y}\|^2}$$

or, for $||\cdot|| = ||\cdot||_2$, $\max_{Y\in\mathbb{R}^{n\times 2}} \dfrac{\operatorname{Var}(\hat{\tau}|X,Y)}{\operatorname{Var}(\hat{\tau}^{\mathrm{CR}}|X,Y)}$.

In particular, if one randomly permutes a single *fixed* partition then

$$\max_{Y\in\mathbb{R}^{n\times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\operatorname{Var}(\hat{\tau}^{CR}|X,Y)} = n-1. \tag{2.4}$$

EXAMPLE 2.1. Fix $n = 2^b$ a power of two and $m = 2$. Let

$$X_i = \sum_{t=0}^{b-\max\{2,\log_2 i\}} (-1)^{\lceil i/2^{t-1}\rceil} 2^{-2^{b-1}+2^{b-t-1}+\left(i-1 \bmod 2^{t-1}\right)},$$
$$Y_i = (-1)^i = (-1)^{\log_2(\operatorname{round}(|X_i|))}.$$

This rather complicated construction essentially yields

$$X \approx \operatorname{round}(X) = \left(-1,-2,-4,\ldots,-2^{2^{b-1}-1},1,2,4,\ldots,2^{2^{b-1}-1}\right)$$

with some perturbations so that the assignment $W = (1,2,1,2,\ldots,1,2)$ uniquely minimizes the group-wise Mahalanobis distance of [7]. Although X_i completely determines Y_{ik}, we are going to see that complete randomization beats blocking, pairwise matching, and re-randomization in this case. For blocking for $b \geq 4$, let us coarsen the space of baseline covariates into eight consecutive intervals so that each contains the same number of subjects, 2^{b-3}. For pairwise matching, let us use the pairwise Mahalanobis distance. And, for re-randomizaiton of [7], we consider both a 1% acceptance probability and an infinitesimal acceptance probability that essentially minimizes the group-wise Mahalanobis metric. We plot the resulting conditional variances $\operatorname{Var}\left(\hat{\tau}\middle|X,Y\right)$ in Figure 1. Specifically, we get that complete randomization has a variance of $4/(n-1)$ whereas blocking has $4/(n-8)$, pairwise matching has $8/n$, and re-randomization with infinitesimal acceptance probability has 4, which realizes the worst-case ratio of (2.4) (it can be verified that this construction also realizes the corresponding worst-case ratios for blocking and pairwise matching). The variance of re-randomization with 1% acceptance is similar to infinitesimal acceptance probability for small n and becomes more similar to randomization as n grows. In each case, complete randomization does better, providing a concrete example of the conclusion of Theorem 2.1.

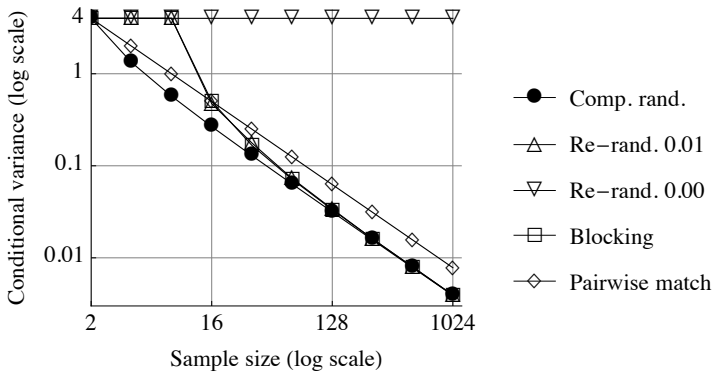

Fig 1: Variance of estimating effect size under various designs in Example 2.1 conditional on the given X and Y values.

2.2. *Structural information and optimal designs.* In the above we argued that from a minimax-variance perspective, complete randomization is optimal when no structural information about the dependence between X_i and Y_{ik} is available. We now consider the effect of such information, which we express as structure on the conditional expectations of outcomes.

Let us denote

$$f_k(x) := \mathbb{E}\left[Y_{ik}\middle|X_i = x\right] \quad \text{and} \quad \epsilon_{ik} := Y_{ik} - f_k\left(X_i\right).$$

The non-random function f_k is interchangeably called the *conditional expectation function* or *regression function*. The law of iterated expectation yields that ϵ_{ik} has mean 0, is mean-independent of X_i, and is uncorrelated with any function of X_i. Combined with independence of subjects, this yields[2]

$$\mathrm{Var}\left(\hat{\tau}\right) = \mathbb{E}\left[\mathrm{Var}\left(B(W,\hat{f})\middle|X\right)\right] + \frac{1}{n}\mathrm{Var}\left(\epsilon_{11}+\epsilon_{12}\right) + \mathrm{Var}\left(\mathrm{SATE}\right),$$

$$\text{where } B(W,\hat{f}) = \frac{2}{n}\sum_{i:W_i=1}\hat{f}(X_i) - \frac{2}{n}\sum_{i:W_i=2}\hat{f}(X_i),\ \ \hat{f}(x) = \frac{f_1(x)+f_2(x)}{2}.$$

As before, the marginal variances of SATE and of $(\epsilon_{11}+\epsilon_{12})$ are completely independent of our choice of design and an efficient design will seek to minimize $\mathrm{Var}\left(B(W,\hat{f})\middle|X\right) = \mathbb{E}\left[B(W,\hat{f})^2\middle|X\right]$ path-by-path, i.e. for the given subjects. Now the unknown is $\hat{f}$ and we let Nature choose it adversarially.

[2]Theorem 3.1 gives an explicit derivation of this decomposition (for general $m \geq 2$).

We will seek to minimize $\mathbb{E}\left[B(W,\hat{f})^2\middle|X\right]$ relative to the magnitude of $\hat{f}$, instead of the magnitude of $\hat{Y}$.

To define a magnitude of $\hat{f}$, we assume that $f_k \in \mathcal{F}\ \forall k$, where $\mathcal{F}$ is a normed vector space with norm $||\cdot|| : \mathcal{F} \to \mathbb{R}_+$. This will represent our structural information about the dependence between X_i and Y_{ik}. This space is a subspace of the vector space $\mathcal{V}$ of all functions $\mathcal{X} \to \mathbb{R}$ under the usual point-wise addition and scaling. For functions f that are not in $\mathcal{F}$ we formally define $||f|| = \infty$. When $\mathcal{F}$ is finite-dimensional, the assumption $\|f_k\| < \infty$ is a parametric one. When $\mathcal{F}$ is infinite-dimensional, it is non-parametric.

Because $B(W,\hat{f})$ is invariant to constant shifts to $\hat{f}$, i.e.,

$$B(W,\hat{f}) = B(W,\hat{f}+c) \quad \text{for } c \in \mathbb{R} \text{ representing a constant function } x \mapsto c,$$

we will want to factor this artifact away. The quotient space $\mathcal{F}/\mathbb{R}$ consists of the classes $[f] = \{f + c : c \in \mathbb{R}\}$ with the norm $||[f]|| = \min_{c\in\mathbb{R}} ||f + c||$. Without loss of generality, we always restrict to this quotient space and write $||f||$ to actually mean the norm in this quotient space. Moreover, for worst-case variances to exist, we will restrict our attention to Banach spaces and require that differences in evaluations are continuous (i.e., the map $f \mapsto (f(X_i) - f(X_j))$ is continuous for each i, j). A Banach space is a normed vector space that is a complete metric space (see [17] and Chapter 10 of [18]).

With all structural information summarized by $\|f_k\| < \infty$, the motivation for the designs we develop next is the bound on the variance that arises:

$$\mathbb{E}\left[B^2(W,\hat{f})|X\right] \le \|\hat{f}\|^2 \max_{f\in\mathcal{F}} \frac{\mathbb{E}\left[B^2(W,f)|X\right]}{\|f\|^2} = \|\hat{f}\|^2 \max_{\|f\|\le 1} \mathbb{E}\left[B^2(W,f)|X\right].$$

Minimizing the above bound is independent of the actual value of $\|\hat{f}\|$ as it merely scales the objective. We will study this bound further and in greater generality in Theorems 3.1 and 3.2, leaving this as mere motivation for now.[3]

Borrowing terminology from game theory, we define two type of designs that seek to minimize this bound: the *pure-strategy optimal design* and the *mixed-strategy optimal design*. We now consider general $m \ge 2$ and define

$$B_{kk'}(W,\hat{f}) = \frac{1}{p}\sum_{i:W_i=k} \hat{f}(X_i) - \frac{1}{p}\sum_{i:W_i=k'} \hat{f}(X_i).$$

The pure-strategy optimal design finds single assignments W that on their own minimize these quantities.

[3]It can also be noted that this bound is of the same form as the objective in Theorem 2.1 but employing the potentially non-symmetric norm $\|\hat{Y}\| = \min_{f(X_i)=\hat{Y}_i} ||f||$ induced by the quotient of $\mathcal{F}$ over the subspace $\{f \in \mathcal{F} : f(X_i) = 0\ \forall i\}$.

DEFINITION 2.1. Given subjects' baseline covariates $X \in \mathcal{X}^n$ and a magnitude function $||\cdot|| : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$, the *pure-strategy optimal design* chooses W uniformly at random from the set of optimizers

$$W \in \arg\min_{W \in \mathcal{W}} \left\{ M^2_{\text{p}}(W) := \max_{||f|| \leq 1} \max_{k \neq k'} B^2_{kk'}(W, f) \right\}.$$

We denote by $M^2_{\text{p-opt}}$ the random variable equal to the optimal value.

The mixed-strategy optimal design directly optimizes the distribution of assignments.

DEFINITION 2.2. Given subjects' baseline covariates $X \in \mathcal{X}^n$ and a magnitude function $||\cdot|| : \mathcal{V} \to \mathbb{R} \cup \{\infty\}$, the *mixed-strategy optimal design* draws W randomly according to a distribution σ such that

$$\sigma \in \arg\min_{\sigma \in \Delta} \left\{ M^2_{\text{m}}(\sigma) := \max_{||f|| \leq 1} \max_{k \neq k'} \sum_{W \in \mathcal{W}} \sigma(W) B^2_{kk'}(W, f) \right\}.$$

We denote by $M^2_{\text{m-opt}}$ the random variable equal to the optimal value.

Both designs satisfy (2.1)-(2.3). The pure-strategy optimal design does due to the symmetry of the objective function (thus, if W is optimal then a treatment-permutation of W is also optimal). The mixed-strategy optimal design does by the construction of Δ. Because the pure-strategy optimal design is feasible in Δ, it is also immediate that $M^2_{\text{m-opt}} \leq M^2_{\text{p-opt}}$.

The objectives $M^2_{\text{p}}(W)$ and $M^2_{\text{m}}(\sigma)$ are the imbalance metrics that the designs seek to minimize. The two are different in nature as one expresses imbalance of a single assignment and the other the imbalance of a whole design. Since evaluation differences are linear and by assumption continuous, both $M^2_{\text{p}}(W)$ and $M^2_{\text{m}}(\sigma)$ are in fact norms taken in the continuous dual Banach space (and this guarantees they are defined). For mixed strategies, $M^2_{\text{m}}(\sigma)$ is actually determined by $n(n-1)/2$ sufficient statistics from σ.

THEOREM 2.2. *Let $\sigma \in \Delta$ be given. Then*

$$M^2_m(\sigma) = M^2_m(P(\sigma)) := \max_{||f|| \leq 1} \frac{2}{pn} \sum_{i,j=1}^{n} P_{ij}(\sigma) f(X_i) f(X_j),$$

where $P_{ij}(\sigma) = \sigma(\{W_i = W_j\}) - \frac{1}{m-1}\sigma(\{W_i \neq W_j\})$.

In the case of $m = 2$, $P(\Delta) = \mathcal{P}$ is the space of feasible P matrices, which are always positive semi-definite (i.e., symmetric with nonnegative eigenvalues).

2.3. *Structural information and existing designs and imbalance metrics.* We now show how the above framework of optimal design in fact recovers various existing designs that balance baseline covariates a priori. In this section we consider two treatments, $m = 2$.

2.3.1. *Blocking and complete randomization.* Randomized block designs are probably the most common non-completely-randomized designs. In a complete block design the sample is segmented into b disjoint evenly-sized blocks $\{i_{1,1}, \dots, i_{1,2p_1}\}, \dots, \{i_{b,1}, \dots, i_{b,2p_b}\}$ so that baseline covariates are equal within each block and unequal between blocks, i.e., $X_{i_{\ell,j}} = X_{i_{\ell',j'}}$ if and only if $\ell = \ell'$. (If any coarsening is done, we assume it was done prior and X_i represents the coarsened value.) Then complete randomization is applied to each block separately and independently of the other blocks.

A complete block design is not always feasible, e.g. when there are subjects with a unique value of covariates or there is an otherwise odd number of subjects with a particular equal value of covariates. In an incomplete block design, there are left-over subjects $i_{0,1}, \dots, i_{0,b'}$. One blocks subjects into evenly-sized blocks so that the number b' is as small as possible, breaking ties randomly as to which subject is left over; complete randomization is then also applied to the left-overs.[4]

Complete blocking can be thought of as minimizing a binary measure of imbalance: 0 if the sets of baseline covariates in each experimental groups are *exactly* the same, infinity otherwise. Incomplete blocking can be thought of as minimizing a discrete measure of imbalance equal to the complement of the number of exact perfect matches across experimental groups (i.e., b'). If complete blocking is feasible, then incomplete blocking necessarily recovers it. If all values of X_i are distinct, then incomplete blocking is the same as complete randomization. As it is the most general, we will only treat incomplete blocking. It turns out that incomplete blocking's exact matching metric corresponds to the space L^∞, i.e., the space $\mathcal{F}$ of bounded functions endowed with the norm $||f||_\infty = \sup_{w \in W} f(w)$.

THEOREM 2.3. *Let $||f|| = ||f||_\infty$. Then the pure-strategy optimal design is equivalent to incomplete blocking.*

As noted before, this also recovers complete blocking (if it is feasible) and complete randomization (if all subjects' baseline covariates are distinct).

[4]Incomplete block designs are much more general than this and cover a much larger scope, especially when treatments outnumber block size, but in our simple setup they amount to breaking ties randomly while maintaing an even partition.

2.3.2. *Pairwise matching.* In optimal pairwise matching, two treatments are considered, subjects are put into pairs so to minimize the sum of pairwise distances in their covariates, and then each pair is split randomly among the two treatments. Any pairwise distance metric δ on $\mathcal{X}$ can be chosen to define the pairwise distances $\delta(X_i, X_j)$. Usually the pairwise Mahalanobis distance is used for vector-valued covariates. The motivation behind pairwise matching is that subjects with similar covariates should have similar outcomes. This corresponds to the space of Lipschitz functions.

THEOREM 2.4. *Let a distance metric δ on $\mathcal{X}$ be given. Let*

$$||f|| = ||f||_{lip} = \sup_{x \neq x'} \frac{f(x) - f(x')}{\delta(x, x')}.$$

Then the pure-strategy optimal design is equivalent to optimal pairwise matching with respect to the pairwise distance metric δ.[5]

COROLLARY 2.1. *Let $\delta_0 > 0$ and a distance metric δ be given. Define $\delta'(x, x') = \max\{\delta(x, x'), \delta_0\}$ for $x \neq x'$ and $\delta'(x, x) = 0$. Let $||f|| = ||f||_{lip}$ with respect to δ'. Then the pure-strategy optimal design is equivalent to caliper matching if it is feasible, i.e., choose at random from pairwise matchings that have all pairwise distances at most δ_0 after blocking exact matches.*

This interpretation of pairwise matching recasts its motivation as structure. Comparing with blocking we see that, whereas blocking treats any two subjects with unequal covariates as potentially having expected outcomes that are as different as any, pairwise matching presumes that unequal but similar covariates should lead to similar expected outcomes. This interpretation of pairwise matching also allows us to generalize it to $m \geq 3$ by using the same space of Lipschitz functions and employing our definition of the optimal designs for general m. We study these new designs in Section 4.1.2.

If we modify the norm and augment it with the sup-norm, we will instead recover an a priori (rather than on-the-fly) version of the method of [10].

THEOREM 2.5. *Let $\delta_0 > 0$ and a distance metric δ be given and let*

$$||f|| = \max\left\{||f||_{lip}, ||f||_\infty / \delta_0\right\}.$$

[5]While $||\cdot||_{\text{lip}}$ is only a seminorm on functions (i.e., $||f||_{\text{lip}} = 0$ doesn't necessarily mean $f = 0$), in the quotient space with respect to constant functions (the kernel of this seminorm) it is a norm and it forms a Banach space. Evaluation differences are well-defined and continuous because they are bounded, $|f(X_i) - f(X_j)| \leq ||f||_{\text{lip}} \delta(X_i, X_j)$.

Then the pure-strategy optimal design is equivalent to the following: minimizes the sum of pairwise distances with respect to δ with the option of leaving a subject unmatched at a penalty of δ_0 (thus no pairs at a distance greater than $2\delta_0$ will ever be matched); then matched pairs are randomly split between the two groups and unmatched subjects are completely randomized.

2.3.3. *Re-randomization of [7].* The method of [7] formalizes the common, but arguably often haphazard, practice of re-randomization as a principled, theoretically-grounded a priori balancing method. The authors consider two treatments, vector-valued baseline covariates $\mathcal{X} = \mathbb{R}^d$, and an imbalance metric equal to a group-wise Mahalanobis metric

$$M^2_{[7]}(W) = \left(\frac{2}{n}\sum_{i=1}^n u_i X_i\right)^T \hat{\Sigma}^{-1} \left(\frac{2}{n}\sum_{i=1}^n u_i X_i\right), \tag{2.5}$$

where $\hat{\Sigma}$ is the sample covariance matrix of X. The authors reinterpret re-randomization as a heuristic algorithm that repeatedly draws random W in order to solve the constraint satisfaction problem $\exists ?W : M^2_{[7]} \leq t$ for a given t (they also propose a normal-approximation method for selecting t to correspond to a particular acceptance probability of a random W).

We can recover (2.5) using our framework. Let $\mathcal{F} = \operatorname{span}\{1, x_1, \ldots, x_d\}$ and define $||f||^2 = \beta^T \hat{\Sigma} \beta + \beta_0^2$ for $f(x) = \beta_0 + \beta^T x$. Using duality of norms,

$$M^2_{\mathrm{p}}(W) = \max_{||f||\leq 1} B^2(W, f) = \left(\max_{\beta^T \hat{\Sigma} \beta \leq 1} \beta^T \left(\frac{n}{2}\sum_{i=1}^n u_i X_i\right)\right)^2 = M^2_{[7]}(W).$$

In [7], the authors argue that when a linear model is known to hold, i.e.,

$$Y_{ik} = \beta_0 + \beta^T X_i + \tau \mathbb{I}[k=1] + \epsilon_i \qquad i = 1, \ldots, n,\ k = 1, 2, \tag{2.6}$$

then fixing t and re-randomizing until $M^2_{[7]}(W) \leq t$ yields a reduction in variance relative to complete randomization that is constant over n:

$$1 - \operatorname{Var}(\hat{\tau}) / \operatorname{Var}(\hat{\tau}^{\mathrm{CR}}) = \eta(1 - \operatorname{Var}(\epsilon_i) / \operatorname{Var}(Y_{i1})),\ \eta \in (0,1) \text{ constant over } n.$$

For us, the imbalance metric is a direct consequence of structure ((2.6) implies $f_k \in \mathcal{F}$) and fully minimizing $M^2_{\mathrm{p}}(W)$ leads to near-best-possible reduction in variance (see Corollary 3.1 and Section 3.3):

$$1 - \operatorname{Var}(\hat{\tau}) / \operatorname{Var}(\hat{\tau}^{\mathrm{CR}}) \longrightarrow 1 - \operatorname{Var}(\epsilon_i) / \operatorname{Var}(Y_{i1}) \quad \text{at a linear rate } 2^{-\Omega(n)}.$$

It is important to keep in mind, however, that the assumption that such a finite-dimensional linear model (2.6) is valid is a parametric, and therefore fragile, assumption. Indeed, we saw in Example 2.1 that fully minimizing $M^2_{[7]}$ when the model is misspecified can lead to worse variance.

2.3.4. *Other finite-dimensional spaces and the method of [13].* We can generalize the idea of parametric balancing methods using finite-dimensional spaces with general norms. Consider any finite-dimensional subspace of $\mathcal{V}$, $\mathcal{F} = \operatorname{span}\{\phi_1, \ldots, \phi_r\}$, and any norm on it. Any such space is always a Banach space and evaluations are always continuous (see Theorems 5.33 and 5.35 of [19]). An important example is the q-norm: $\|\beta_1\phi_1 + \cdots + \beta_r\phi_r\| = ||\beta||_q$ where $||\beta||_q = (\sum_i |\beta_i|^q)^{1/q}$ for $1 \le q < \infty$ and $||\beta||_\infty = \max_i |\beta_i|$. This yields

$$M_{\mathrm{p}}^2(W) = \left|\left|\left(\frac{n}{2}\sum_{i=1}^{n} u_i\phi_1(X_i), \ldots, \frac{n}{2}\sum_{i=1}^{n} u_i\phi_r(X_i)\right)\right|\right|_{q^*}^2$$

for $1/q + 1/q^* = 1$. Hence, the optimal design matches the sample ϕ_j moments between the groups by minimizing a norm in the vector of mismatches.

The covariance-scaled 2-norm on $\mathcal{F} = \operatorname{span}\{1, x_1, \ldots, x_d\}$ was considered in Section 2.3.3 and gave rise to the group-wise Mahalanobis metric. Endowing $\mathcal{F} = \operatorname{span}\{1, x_1, \ldots, x_d, x_1^2/\rho, \ldots, x_d^2/\rho, x_1x_2/(2\rho), \ldots, x_{d-1}x_d/(2\rho)\}$ with the ∞-norm and normalizing the data will recover the method of [13].

2.4. *New designs using RKHS structure.* In our framework, one starts with structural information about the relationship between X_i and Y_{ik} and this leads to measures of imbalance and to optimal designs that minimize them. In the previous section we saw how different structures led to well-known measures of imbalance and designs. We now explore how other choices of structure lead to new designs. We treat general $m \ge 2$ in this section.

We will express structure using reproducing kernel Hilbert spaces (RKHS). A Hilbert space is an inner-product space such that the norm induced by the inner product, $\|f\|^2 = \langle f, f\rangle$, yields a Banach space. An RKHS $\mathcal{F}$ is a Hilbert space of functions for which evaluation $f \mapsto f(x)$ is continuous for each $x \in \mathcal{X}$ (see [20]). Continuity and the Riesz representation theorem imply that for each $x \in \mathcal{X}$ there is $\mathcal{K}(x, \cdot) \in \mathcal{F}$ such that $\langle \mathcal{K}(x, \cdot), f(\cdot)\rangle = f(x)$ for every $f \in \mathcal{F}$. The symmetric map $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called the reproducing kernel of $\mathcal{F}$. The name is motivated by the fact that $\mathcal{F} = \operatorname{closure}_{\mathcal{F}}(\operatorname{span}\{\mathcal{K}(x, \cdot) : x \in \mathcal{X}\})$. Thus $\mathcal{K}$ fully characterizes $\mathcal{F}$. Prominent examples of kernels are:

1. The linear kernel $\mathcal{K}(x, x') = x^T x'$. This spans the finite-dimensional space of linear functions and induces a 2-norm on coefficients.
2. The polynomial kernel $\mathcal{K}_s(x, x') = (1 + x^T x'/s)^s$. It spans the finite-dimensional space of all polynomials of degree up to s.
3. Any kernel $\mathcal{K}(x, x') = \sum_{i=0}^{\infty} a_i (x^T x')^i$ with $a_i \ge 0$ (subject to convergence). This includes the previous two examples. Another case is the

exponential kernel $\mathcal{K}(x,x') = e^{x^T x'}$, which can be seen as the infinite-dimensional limit of the polynomial kernel. The corresponding space is infinite-dimensional (non-parametric).

4. The Gaussian kernel $\mathcal{K}(x,x') = e^{-||x-x'||^2}$. The corresponding space is infinite-dimensional (non-parametric) and is studied in [21].

For given $X \in \mathcal{X}^n$ and an RKHS with kernel $\mathcal{K}$, we will often use the Gram matrix $K_{ij} = \mathcal{K}(X_i, X_j)$. The Gram matrix is always positive semi-definite and as such it has a matrix square root $K = \sqrt{K}\sqrt{K}$.

As mentioned above, an RKHS induces a norm. Therefore, in our framework, it also induces imbalance metrics and optimal designs.

THEOREM 2.6. *Let $\mathcal{F}$ be an RKHS with kernel $\mathcal{K}$. Then,*

$$M_p^2(W) = \frac{1}{p^2} \max_{k \neq k'} \sum_{i,j=1}^{n} (w_{ik} - w_{ik'}) K_{ij} (w_{jk} - w_{jk'}), \quad \textit{and} \tag{2.7}$$

$$M_m^2(P) = \frac{2}{np} \lambda_{\max} \left(\sqrt{K} P \sqrt{K} \right). \tag{2.8}$$

Notice that (2.7) corresponds to a discrepancy statistic known as *maximum mean discrepancy* between the experimental groups. Maximum mean discrepancy is used as a test statistic in two-sample testing (see [22, 23, 24]).

The problem of minimizing (2.7) or (2.8) can be interpreted as a multi-way multi-criterion number partitioning problem. For $m = 2$, $\mathcal{X} = \mathbb{R}$, and $\mathcal{K}(x,x') = xx'$ ($K = XX^T$), we get the usual balanced number partitioning problem for both (2.7) and (2.8): recalling our definitions of $\mathcal{U}$ and $\mathcal{P}$,

$$\frac{n}{2} M_{\text{p-opt}} = \sqrt{\min_{u \in \mathcal{U}} u^T (XX^T) u} = \min_{u \in \mathcal{U}} \left| \sum_{i=1}^{n} u_i X_i \right|,$$

$$\frac{n}{2} M_{\text{m-opt}} = \sqrt{\min_{P \in \mathcal{P}} \operatorname{trace}\left(P(XX^T)\right)} = \min_{u \in \mathcal{U}} \left| \sum_{i=1}^{n} u_i X_i \right|,$$

where the last equality is due to the facts that $\lambda_{\max}(M) = \operatorname{trace}(M)$ if M is rank-1 positive semi-definite and that a linear objective on a polytope is optimized at a corner point. This reduction also shows that both problems are NP-hard (see problem [SP12] and comment on p. 223 of [25]).

Such partitioning problems generically have unique optima up to permutation so the pure-strategy optimal design usually randomizes among the $m!$ permutations of a single partition of subjects. This is not generally the case for the mixed-strategy optimal design. Consider $m = 2$. Since the affine

hull of $\mathcal{U}$ is $(n-1)$-dimensional, the mixed-strategy optimal design mixes at the very least $2(\operatorname{rank}(K)-1)$ assignments. Moreover, by Carathéodory's theorem any $P \in \mathcal{P}$ can be identified as the convex combination of $n(n-1)$ points in $\{uu^T : u \in \mathcal{U}\}$ (whose affine hull is $(n(n-1)-1)$-dimensional) so that the mixed-strategy objective $M_{\mathrm{m}}^2(\sigma)$ of any a priori balancing design $\sigma \in \Delta$ can also be achieved by mixing no more than $2n(n-1)$ assignments.

In Sections 4.1.3 and 4.2 we will study how we solve the pure- and mixed-strategy optimal designs, respectively. For now let us consider two concrete examples with the various designs we have so far studied.

EXAMPLE 2.2. Consider the following setup: we measure $d \geq 2$ baseline covariates for each subject that are uniformly distributed in the population $X_i \sim \operatorname{Unif}\left([-1,1]^d\right)$, the two treatments $m = 2$ have constant individual effects $Y_{i1} - Y_{i2} = \tau$, and the conditional expectation of outcomes depends on two covariates only $\mathbb{E}\left[Y_{i1} \middle| X = x\right] - \tau/2 = \mathbb{E}\left[Y_{i2} \middle| X = x\right] + \tau/2 = \hat{f}(x_1, x_2)$. We consider a variety of conditional expectation functions:[6]

Linear: $\hat{f}(x_1, x_2) = x_1 - x_2$.
Quadratic: $\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2$.
Cubic: $\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2 + x_1^3 - x_2^3 - 3x_1^2x_2 + 3x_1x_2^3$.
Sinusoidal: $\hat{f}(x_1, x_2) = \sin(\frac{\pi}{3} + \frac{\pi x_1}{3} - \frac{2\pi x_2}{3}) - 6\sin(\frac{\pi x_1}{3} + \frac{\pi x_2}{4}) + 6\sin(\frac{\pi x_1}{3} + \frac{\pi x_2}{6})$.

To simulate the common situation where some covariates matter and some do not and which is which is not known a priori, we consider both the case $d = 2$ (only balance the relevant covariates) and $d = 4$ (also balance some covariates that turn out to be irrelevant).

We consider the following designs: (1) complete randomization, i.e., the pure-strategy optimal design for L^∞; (2) blocking on the orthant of X_i (d two-level factors), i.e., the pure-strategy optimal design for L^∞ after coarsening; (3) re-randomization with 1% acceptance probability and Mahalanobis objective; (4) pairwise matching with Mahalanobis distance, i.e., the pure-strategy optimal design for the Lipschitz norm; (5) the pure-strategy optimal design with respect to the linear kernel; (6) the pure-strategy optimal design with respect to the quadratic kernel (polynomial kernel with $s = 2$); (7) the mixed-strategy optimal design with respect to the Gaussian kernel; and (8) the mixed-strategy optimal design with respect to the exponential kernel.[7] All of these designs result in an unbiased estimate of SATE = PATE = τ and can therefore be compared on their variance.

[6] We do not consider the case of no relationship ($\hat{f}(x_1, x_2) = c$) because Theorem 3.1 proves that in this case any a priori balancing design yields the same estimation variance.

[7] For the mixed-strategy designs we use the heuristic solution given by Algorithm 4.3.

In Figure 2 we plot the variances of the resulting estimators relative to $V_n = \mathrm{Var}\left(\mathrm{SATE}\right) + \mathrm{Var}\left(\epsilon_{11} + \epsilon_{12}\right)/n$ (see Theorem 3.1).

There are several features to note. One is that when a parametric model is correctly specified and specifically optimized for, the variance (relative to V_n) shrinks linearly (inverse exponentially) – we argue this is a general phenomenon in Section 3.3. This phenomenon is clearest in the case of linear conditional expectation and the pure-strategy optimal design with respect to the linear kernel, but the same design does not do so well when the linear model is misspecified. The pure-strategy optimal design with respect to the quadratic kernel also has a linear, but slower, convergence for the linear conditional expectation, but it performs better in the other cases, both when a quadratic model is correctly specified and when it is not. The mixed-strategy optimal designs with respect to the Gaussian and exponential kernels seem to have uniformly good performance in all cases and in particular still exhibit what would seem to be linear convergence for the linear and quadratic cases.[8] It would seem that these non-parametric methods strike a good compromise between efficiency and robustness. Finally, we note that compared to balancing only those covariates that matter most ($d = 2$), balancing also other covariates ($d = 4$) leads to loss of efficiency, as would be expected, but the order of convergence (linear) is the same.

EXAMPLE 2.3. We now consider the effect of a priori balance on a real dataset. We use the diabetes study dataset from [26] described therein as follows: "Ten [$d = 10$] baseline variables [X_i], age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of [442] diabetes patients, as well as the response of interest [Y_i'], a quantitative measure of disease progression one year after baseline." We consider a hypothetical experiment where the prognostic features X_i are measured at the onset, a control or treatment is applied, and the response after one year is measured. In our hypothetical setup, the treatment reduces disease progression by exactly τ so that $Y_{i1} = Y_i'$ and $Y_{i2} = Y_i' - \tau$. Fixing n, we draw n subjects with replacement from the population of 442, normalize the covariate data so that the sample of n has zero sample mean and identity sample covariance and divide by $d = 10$, apply each of the a priori balancing designs considered in Example 2.2 to the normalized covariates, and finally apply the treatments and measure the responses and the mean differences $\hat{\tau}$. Again, we consider either balancing all $d = 10$ covariates or only the

[8]The argument in Section 3.3 concerns only finite-dimensional spaces and does not support this observation as a general phenomenon.

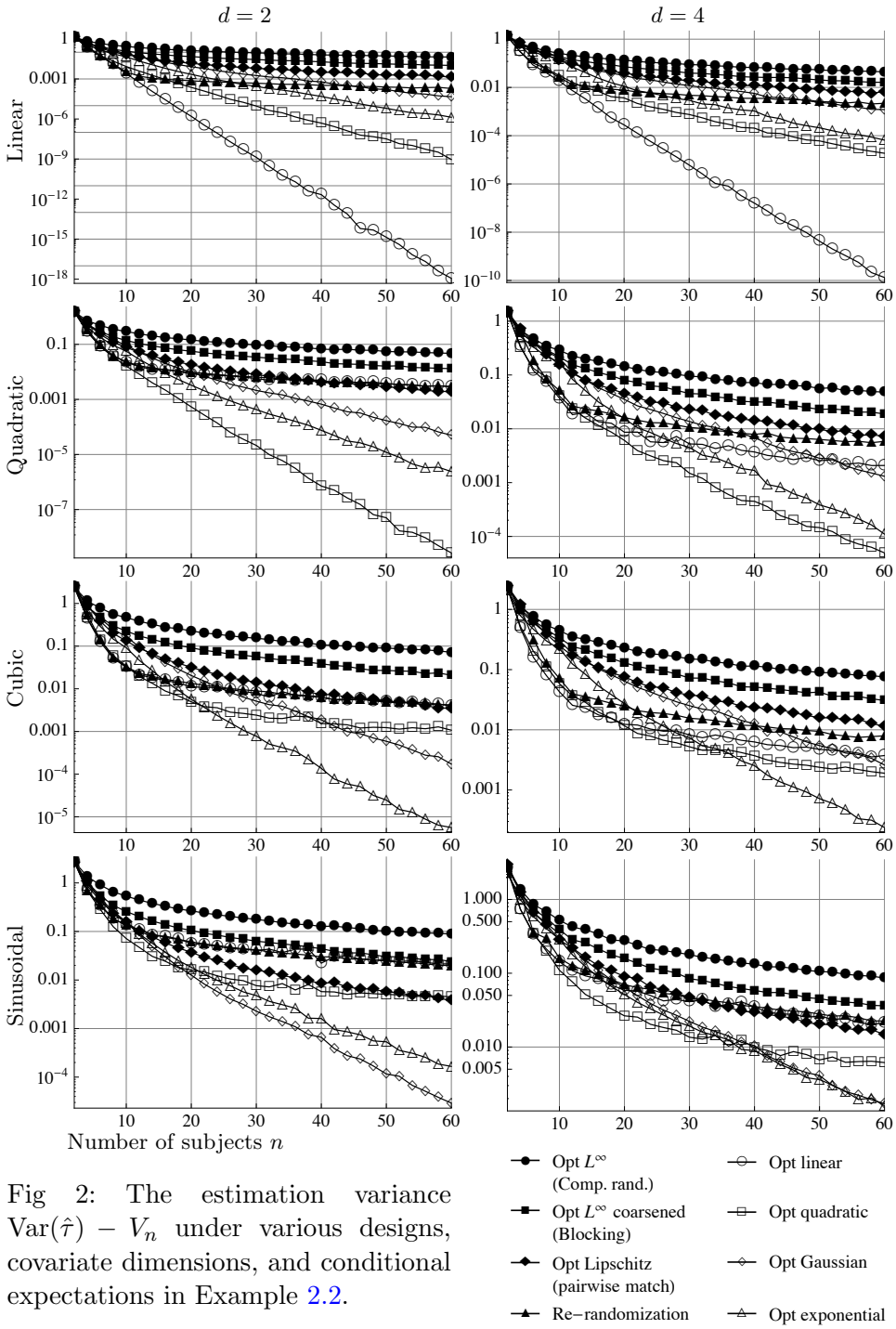

Fig 2: The estimation variance $\mathrm{Var}(\hat{\tau}) - V_n$ under various designs, covariate dimensions, and conditional expectations in Example 2.2.

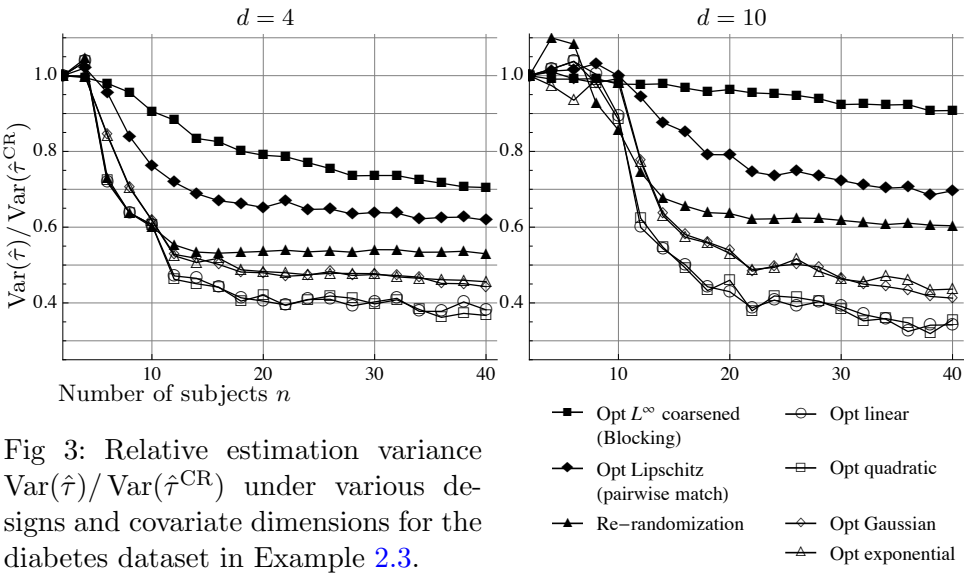

Fig 3: Relative estimation variance $\mathrm{Var}(\hat{\tau})/\mathrm{Var}(\hat{\tau}^{\mathrm{CR}})$ under various designs and covariate dimensions for the diabetes dataset in Example 2.3.

$d = 4$ covariates that are ranked first by [26] (these are $\{3, 9, 4, 7\}$). We plot estimation variances relative to complete randomization in Figure 3.

For larger n, the relative variance of each method stabilizes around a particular ratio. Each of blocking, pairwise matching, and re-randomization result in a higher ratio when attempting to balance all covariates compared to balancing only the four most important. For example, re-randomization on all 10 covariates gives ∼60% of complete randomization's variance whereas restricting to the important covariates yields ∼53%. On the other hand, the RKHS-based optimal designs yield lower relative variances for both $d = 10$ and $d = 4$, converging slower for $d = 10$ but using the small additional prognostic content of the extra covariates to reduce variance further. For example, the pure-strategy optimal designs with respect to the linear and quadratic kernels both yield ∼40% of complete randomization's variance for $d = 4$ and ∼35% for $d = 10$, taking only slightly longer to get below ∼40% when $d = 10$. This can be attributed to the linear rate at which the optimal designs eliminate imbalances (see Section 3.3). Thus, even if there are some less relevant variables, all are immediately near-perfectly balanced for modest n; the only limiting factors are the residuals (ϵ_{ik}), which, by definition, cannot be controlled for using the covariates X alone (see Corollary 3.1).

2.4.1. *Aside: a Bayesian interpretation.* The pure-strategy optimal design can also be interpreted in a Bayesian perspective as an optimal design.

The interpretation is very similar to the standard Bayesian interpretation of regularized regression using Gaussian processes (see e.g. [27] and §6.2 of [28]). Let $m = 2$ and let $\mathcal{F}$ be a given RKHS with kernel $\mathcal{K}$. Let us assume a Gaussian prior on $\hat{f}$ with covariance operator $\mathcal{K}$, i.e. $\hat{f}(x)$ is Gaussian for every $x \in \mathcal{X}$ and the covariance of $\hat{f}(x)$ and $\hat{f}(x')$ is equal to $\mathcal{K}(x, x')$. Then we have that the Bayes variance risk of a design W is

$$\begin{aligned}\mathbb{E}\left[B^2(W,f)|X,Y,W\right] &= \frac{4}{n^2}\sum_{i,j=1}^{n} u_i u_j \mathbb{E}\left[f(X_i)f(X_j)|X,Y\right] \\ &= \frac{4}{n^2}\sum_{i,j=1}^{n} u_i u_j \mathcal{K}(X_i,X_j) = \frac{4}{n^2}u^T K u = M_{\mathrm{p}}^2(W).\end{aligned}$$

Note however that randomization is not necessary from a standard Bayesian perspective (for futher discussion see [29, 30]) and therefore a Bayesian design may not satisfy (2.1)-(2.2). In contrast, the pure- and mixed-strategy optimal designs both randomize by construction. Moreover, for the mixed-strategy optimal design, it is generally optimal to randomize beyond just random permutations of one partition.

3. Characterizations of a priori balancing designs. We now try to characterize the estimators that arise from pure- and mixed-strategy optimal designs as well as a priori balancing designs in general. We argue the estimator is unbiased and then bound its variance in terms of a priori imbalance – a result that intimately connects imbalance prior to treatment to variance of estimation after treatment. We also discuss consistency and the convergence rate of imbalance (and hence variance).

3.1. *Variance.* We begin by decomposing the variance of any estimator arising from an a priori balancing design, that is, one satisfying (2.1)-(2.3).

THEOREM 3.1. *Suppose (2.1)-(2.3) are satisfied. Then, for all $k \neq k'$,*

(a) $\hat{\tau}_{kk'}$ is conditionally and marginally unbiased, i.e.,

$$\mathbb{E}\left[\hat{\tau}_{kk'}\middle|X,Y\right] = \mathrm{SATE}_{kk'}, \quad \mathbb{E}\left[\hat{\tau}_{kk'}\right] = \mathrm{PATE}_{kk'}\,.$$

(b) $\hat{\tau}_{kk'} = \mathrm{SATE}_{kk'} + D_{kk'} + E_{kk'}$,

$$\text{where} \quad D_{kk'} := \frac{1}{m}\sum_{l\neq k} B_{kl}\left(f_k\right) - \frac{1}{m}\sum_{l\neq k'} B_{k'l}\left(f_{k'}\right),$$

$$E_{kk'} := \frac{1}{n}\sum_{i=1}^{n}\left((mw_{ik}-1)\epsilon_{ik} - (mw_{ik'}-1)\epsilon_{ik'}\right).$$

(c) $\mathrm{SATE}_{kk'}$, $D_{kk'}$, *and* $E_{kk'}$ *are all uncorrelated so that*

$$\begin{aligned}\operatorname{Var}(\hat{\tau}_{kk'}) = \frac{1}{n}\operatorname{Var}(Y_{1k} - Y_{1k'}) + \operatorname{Var}(D_{kk'}) \\ + \frac{1}{n}\operatorname{Var}(\epsilon_{1k} + \epsilon_{1k'}) + \frac{m-2}{n}\left(\operatorname{Var}(\epsilon_{1k}) + \operatorname{Var}(\epsilon_{1k'})\right).\end{aligned}$$

(Note that the last term drops when only two treatments are considered.)

Note that in part (c), every term except for $\operatorname{Var}(D_{kk'})$ is completely unaffected by any a priori balancing. Below we provide a bound on it based on the expected minimal imbalance produced by an optimal design.

THEOREM 3.2. *If the pure- or mixed-strategy optimal design is used,*

$$\operatorname{Var}(D_{kk'}) \leq \frac{(||f_k|| + ||f_{k'}||)^2}{2}\left(1 - \frac{1}{m}\right)\mathbb{E}\left[M_{opt}^2\right], \tag{3.1}$$

where $M_{opt}^2 = M_{p\text{-}opt}^2$ *or* $M_{opt}^2 = M_{m\text{-}opt}^2$, *respectively.*

In (3.1), $(||f_k|| + ||f_{k'}||)^2$ is unknown but constant, merely scaling the bound.

Combining the two theorems we get that when the pure- or mixed-strategy optimal design is used, the variance of our estimator is bounded as follows:

$$\begin{aligned}\operatorname{Var}(\hat{\tau}_{kk'}) \leq \frac{1}{n}\operatorname{Var}(Y_{1k} - Y_{1k'}) + \frac{(||f_k|| + ||f_{k'}||)^2}{2}\left(1 - \frac{1}{m}\right)\mathbb{E}\left[M_{\mathrm{opt}}^2\right] \\ + \frac{1}{n}\operatorname{Var}(\epsilon_{1k} + \epsilon_{1k'}) + \frac{m-2}{n}\left(\operatorname{Var}(\epsilon_{1k}) + \operatorname{Var}(\epsilon_{1k'})\right).\end{aligned}$$

This intimately connects balance prior to treatment and randomization to estimation variance afterward. For example, for pairwise matching this explicitly connects the sum of pair differences before treatment to estimation variance after via the Lipschitz constant of the unknown regression function.

Basic arithmetic with this bound yields the following simplification.

COROLLARY 3.1. *Suppose* $m = 2$ *and that individual effects are constant* $Y_{i1} - Y_{i2} = \tau$. *Denote* $\sigma^2 = \operatorname{Var}(Y_{i1}) = \operatorname{Var}(Y_{i2})$, $\xi^2 = \operatorname{Var}(\epsilon_{i1}) = \operatorname{Var}(\epsilon_{i2})$, *and* $R^2 = 1 - \xi^2/\sigma^2$ *(explained variance fraction). Then, the variance due to the optimal design relative to complete randomization is bounded as follows:*

$$1 - R^2 \leq \frac{\operatorname{Var}(\hat{\tau})}{\operatorname{Var}(\hat{\tau}^{CR})} \leq 1 - R^2 - \frac{n}{16\sigma^2}(||f_k|| + ||f_{k'}||)^2\,\mathbb{E}\left[M_{opt}^2\right].$$

Alternatively, the relative reduction in variance is simply one minus the above. Despite the constant effect assumption, this bound provides important insights. On the one hand, it says that any a priori balancing effort can never do better than $(1 - R^2)$ relative to complete randomization. This makes sense: balancing based on X alone can only help to the extent that it is predictive of outcomes. On the other hand, it says that if $\mathbb{E}\left[M^2_{\text{opt}}\right]$ decays super-logarithmically, i.e. $o(1/n)$, then the relative variance converges to the best possible, which is $(1 - R^2)$. In Section 3.3 we study a case where the convergence is linear, i.e. $2^{-\Omega(n)}$, much faster than logarithmic.

When $f_k \notin \mathcal{F}$ we have $||f_k|| = \infty$ and the bound (3.1) is trivial. Accounting for the distance between f_k and $\mathcal{F}$, an alternative bound is possible.

THEOREM 3.3. *If the pure- or mixed-strategy optimal design is used,*

$$\operatorname{Var}(D_{kk'}) \leq \left(1 - \frac{1}{m}\right) \inf_{g_k, g_{k'} \in \mathcal{F}} \Big(\left(||g_k|| + ||g_{k'}||\right)^2 \mathbb{E}\left[M^2_{opt}\right] + \frac{2}{m} \left(||f_k - g_k||_2 + ||f_{k'} - g_{k'}||_2\right)^2 \Big),$$

where $M^2_{opt} = M^2_{p\text{-}opt}$ or $M^2_{opt} = M^2_{m\text{-}opt}$, respectively, and $||g||_2^2 = \mathbb{E}\left[g(X_1)^2\right]$ is the L^2 norm with respect to the measure of X_1. (By the assumption that potential outcomes have second moments, we have $||f_k||_2 < \infty$.)

3.2. *Consistency.* An estimator is said to be strongly consistent if it converges almost surely to the estimand, the quantity it tries to estimate. In light of Theorem 3.1(b), an a priori balancing design results in a strongly consistent estimator if and only if $D_{kk'}$ converges to 0 almost surely (since $\text{SATE}_{kk'} + E_{kk'}$ is already strongly consistent). Employing laws of large numbers in Banach spaces, we can express sufficient conditions for strong consistency in terms of a functional analytical property of $\mathcal{F}$ known as B-convexity.

DEFINITION 3.1. A Banach space is said to be *B-convex* if there exists $N \in \mathbb{N}$ and $\eta < N$ such that for every $g_1, \ldots, g_N$ with $||g_i|| \leq 1\ \forall i$ there exists a choice of signs so that $||\pm g_1 \pm \cdots \pm g_N|| \leq \eta$.

It is easy to verify that all the Banach spaces so far considered are B-convex with the exception of L^∞. In particular, every Hilbert space or finite-dimensional Banach space is B-convex. We use this condition to characterize consistency in the following.

THEOREM 3.4. *Suppose $f_k, f_{k'} \in \mathcal{F}$. If either*

(a) $\mathcal{F}$ *is B-convex and* $\mathbb{E}\left(\max_{||f||\leq 1}(f(X_1)-f(X_2))\right)^2 < \infty$ *or*

(b) $\mathcal{F}$ *is a Hilbert space and* $\mathbb{E}\left|\max_{||f||\leq 1}(f(X_1)-f(X_2))\right| < \infty$

then the estimator $\hat{\tau}_{kk'}$ *arising from either the pure- or mixed-strategy optimal design is strongly consistent.*

3.3. *Linear rate of convergence for parametric designs.* In Theorem 3.4, we argued that the estimator converges, i.e., it is consistent, but we did not discuss its rate of convergence. In this section, we study the rate of convergence of $\mathbb{E}M^2_{\text{opt}}$ for the pure- and mixed-strategy designs and hence the convergence of the corresponding estimator's variance as per Theorem 3.2. In particular, we now argue that $\mathbb{E}M^2_{\text{opt}} = 2^{-\Omega(n)}$ for the case $m = 2$ and $\mathcal{F}$ finite dimensional (i.e., parametric). We will also study $m \geq 3$ empirically and observe similar convergence.

Let $\phi_1, \ldots, \phi_r$ be a basis for the finite-dimensional $\mathcal{F}$ and $\Phi_{ij} = \phi_j(X_i)$. Because all norms in finite dimensions are equivalent, i.e., $c\,||\cdot||' \leq ||\cdot|| \leq C\,||\cdot||'$ (see Theorem 5.36 of [19]), it follows that any rate of convergence that applies when $\mathcal{F}$ is endowed with the 2-norm ($||\beta_1\phi_1 + \cdots + \beta_r\phi_r|| = ||\beta||_2$) also applies when $\mathcal{F}$ has any given norm. Next note that since $M^2_{\text{m-opt}} \leq M^2_{\text{p-opt}}$, any rate of convergence for $M^2_{\text{p-opt}}$ applies also to $M^2_{\text{m-opt}}$. So, we restrict our attention to pure-strategy optimal designs under the 2-norm.

Our argument is a heuristic one (not a precise proof) and will follow the asymptotic approximation of the configurations W with energies $M^2_{\text{p}}(W)$ as a spin glass following the random energy model (REM) where energies are assumed independent. This approximation is commonly used to study the distributions of the optima of combinatorial optimization problems with random inputs and has been found to be valid asymptotically for partition problems similar to the one we are considering (see [31, 32, 33]).

Let $\Sigma_{ij} = \text{Cov}\,(\phi_i(X_1), \phi_j(X_1))$ and let $\lambda_1, \ldots, \lambda_{r'} > 0$ be its positive eigenvalues where $r' = \text{rank}\,(\Sigma)$. The distribution of $M^2_{\text{p}}(W)$ is the same for any one fixed W. Fix $W_i = (i \mod 2)+1$ ($u_i = (-1)^{i+1}$). By the multivariate central limit theorem we have the following convergence in distribution,

$$\frac{2}{n}\Phi^T u = \left(\frac{2}{n}\sum_{i=1}^{n/2}(\phi_j(X_{2i-1}) - \phi_j(X_{2i}))\right)_{j=1}^{r} \xrightarrow{d} \mathcal{N}(0, 2\Sigma).$$

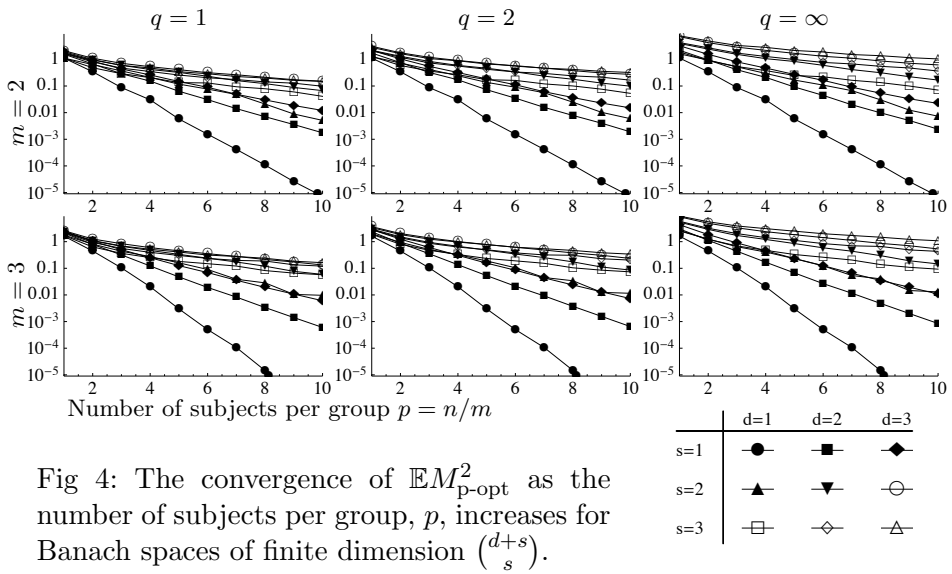

Fig 4: The convergence of $\mathbb{E}M^2_{\text{p-opt}}$ as the number of subjects per group, p, increases for Banach spaces of finite dimension $\binom{d+s}{s}$.

By continuous transformation, we also have

$$M^2_{\text{p}}(W) = \sup_{||\beta||_2 \leq 1} \left(\frac{2}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} u_i \beta_j \phi_j(X_i) \right)^2 = \left|\left| \frac{2}{n} \Phi^T u \right|\right|_2^2 \xrightarrow{d} \sum_{i=1}^{r'} 2\lambda_i \chi^2_1,$$

the weighted sum of independent chi-squared random variables with one degree of freedom. Denote the corresponding CDF by H and PDF by h, which are given in series representation in [34]. In following with the REM approximation we assume independent energies so that $M^2_{\text{p-opt}}$ is distributed as the smallest order statistic among $\binom{n}{n/2}$-many independent draws from H. By Theorem 11.3 of [35] and $\lim_{t \to 0^+} th(t)/H(t) = r'/2$, we have that

$$\mathbb{P}\left(M^2_{\text{p-opt}}/\beta_n \leq t \right) \longrightarrow 1 - \exp\left(-t^{r'/2} \right)$$

for β_n satisfying $H(\beta_n) \cdot \binom{n}{n/2} \to 1$. By formula (40) of [34] this is true for

$$\beta_n = 4 \left(\Gamma\left(r'/2 + 1 \right) \Big/ \binom{n}{n/2} \right)^{2/r'} \prod_{i=1}^{r'} \lambda_i^{1/r'}.$$

Thus, $\mathbb{E}M^2_{\text{p-opt}} \approx \beta_n \Gamma\left(2/r' + 1 \right)$ asymptotically. By Stirling's formula,

$$\mathbb{E}M^2_{\text{m-opt}} \leq \mathbb{E}M^2_{\text{p-opt}} = O\left(2^{-2n/r'} n^{1/r'} \right) = 2^{-\Omega(n)}.$$

We plot the convergence of $\mathbb{E}M^2_{\text{p-opt}}$ for a range of cases in Figure 4. We consider $m = 2, 3$, $X_i \sim \mathcal{N}(0, I_d)$, $\phi_\theta(x) = s^{1-\sum_i \theta_i} \prod_{i=1}^d x_i^{\theta_i}$, $d = 1, 2, 3$, $r = \binom{d+s}{s}$ (all monomials up to degree s) for $s = 1, 2, 3$, and q-norms 1, 2, and ∞. All exhibit linear convergence (note log scale).

4. Algorithms for optimal design. We now address how to actually realize the optimal designs, i.e., solve the optimization problems in the definitions of the pure- and mixed-strategy optimal designs. For complete randomization, blocking, and pairwise matching (with two treatments), how to do so is already clear; here we address the other designs that arouse from our framework. For the pure-strategy optimal designs, the optimization problems will be linear, quadratic, and second-order cone optimization problems subject to integer constraints on some of the variables. Therefore, for these we can use integer optimization software to find the optimal design. In all numerical results in this paper, we use Gurobi v5.6 [36]. For the mixed-strategy optimal design, the problem is too hard to solve exactly and we provide heuristics based on semi-definite optimization.

4.1. *Optimizing pure strategies.* The pure-strategy optimization problem can be written as

$$\begin{aligned}
\sqrt{\min_{W\in\mathcal{W}} M^2_{\text{p}}(W)} = \min_{\lambda\in\mathbb{R},\, w\in\{0,1\}^n} \quad & \lambda \\
\text{s.t. } \lambda \geq \max_{||f||\leq 1} \frac{1}{p} \sum_{i=1}^n (w_{ik} - w_{ik'})\, f(X_i) \quad & \forall k < k' \qquad (4.1)\\
\sum_{k=1}^m w_{ik} = 1 \quad & \forall i = 1, \ldots, n \\
\sum_{i=1}^n w_{ik} = p \quad & \forall k = 1, \ldots, m,
\end{aligned}$$

where we have used the fact that optimizing the square is the same as optimizing the absolute value and then used the symmetry of the norm to remove the absolute value and rid of excess constraints ($k > k'$). What remains is to write the constraints (4.1) in a way fitting for a linear, quadratic, or second-order cone optimization problem. We assume the solver software will arbitrarily return any one optimal solution at random. In case this is not so, we still randomly permute the result to ensure condition (2.2) holds.

4.1.1. *Finite-dimensional q-space.* For the setup as in Section 2.3.4,

$$\max_{||f||\leq 1} \frac{1}{p} \sum_{i=1}^n (w_{ik} - w_{ik'})\, f(X_i)$$

$$= \left|\left| \left(\frac{n}{2} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \phi_1(X_i), \ldots, \frac{n}{2} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \phi_r(X_i) \right) \right|\right|_{q^*}$$

for $1/q + 1/q^* = 1$. It follows that for $q = 1, \infty$, the pure-strategy optimization problem is a linear optimization problem with integer variables. For $q = 2$, the problem for $m = 2$ is a quadratic optimization problem with integer variables and for $m \geq 3$ it is a second-order cone optimization problem with integer variables (the difference being whether the quadratic term is in the objective or constraints). Rational q can also be dealt with using second-order cone optimization via the results of [37]. For example, $m = 2$, $q = 2$, and $\Phi_{ij} = \phi_j(X_i)$, leads to a binary quadratic optimization problem:

$$M^2(W) = \frac{4}{n^2} \min_{u \in \mathcal{U}} u^T \Phi \Phi^T u.$$

4.1.2. *Lipschitz functions.* Given a pairwise distance metric δ, we define the norm $||f|| = ||f||_{\text{lip}}$. When $m = 2$, Theorem 2.4 shows that the pure-strategy optimal design is equivalent to pairwise matching. The corresponding optimization problem is weighted non-bipartite matching, which can be solved in polynomial time using Edmond's algorithm [38]. For $m \geq 3$, we let $D_{ij} = \delta(X_i, X_j)$ and use linear optimization duality [39] to write

$$\lambda \geq \max_{||f|| \leq 1} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) = \max_{ve^T - ev^T \leq D} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) v_i$$

$$\iff \exists S \in \mathbb{R}_+^{n \times n} \text{ s.t. } \begin{array}{l} \lambda \geq \operatorname{trace}(DS)/p, \\ \sum_{j=1}^{n} (S_{ij} - S_{ji}) = w_{ik} - w_{ik'} \quad \forall i = 1, \ldots, n, \end{array}$$

yielding a linear optimization problem with integer variables.

For the modification $||f|| = \max\left\{ ||f||_{\text{lip}}, ||f||_\infty / \delta_0 \right\}$ considered in Theorem 2.5, we can instead write

$$\lambda \geq \max_{||f|| \leq 1} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) = \max_{ve^T - ev^T \leq D, \, ||v||_\infty \leq \delta_0} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) v_i$$

$$\iff \exists \begin{array}{l} S \in \mathbb{R}_+^{n \times n} \\ t \in \mathbb{R}^n \end{array} \text{ s.t. } \begin{array}{l} \lambda \geq (\operatorname{trace}(DS) + \delta_0 ||t||_1)/p, \\ \sum_{j=1}^{n} (S_{ij} - S_{ji}) + t_i = w_{ik} - w_{ik'} \quad \forall i = 1, \ldots, n. \end{array}$$

This also leads to a linear optimization problem with integer variables.

4.1.3. *RKHS.* As in Theorem 2.6 we have

$$\left(\max_{||f|| \leq 1} \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) \right)^2 = \frac{1}{p} \sum_{i,j=1}^{n} (w_{ik} - w_{ik'}) K_{ij} (w_{jk} - w_{jk'}).$$

Therefore, for $m = 2$ the pure-strategy optimization problem is a quadratic optimization problem with integer variables and for $m \geq 3$ it is a second-order cone optimization problem with integer variables. Namely, for $m = 2$, we get the binary quadratic optimization problem:

$$M^2(W) = \frac{4}{n^2} \min_{u \in \mathcal{U}} u^T K u.$$

4.2. *Optimizing mixed strategies.* For the case of mixed strategies we only consider the case of $m = 2$ and $\mathcal{F}$ being an RKHS. As per Theorems 2.2 and 2.6, the corresponding optimization problem is

$$\frac{4}{n^2} \min_{P \in \mathcal{P}} \lambda_{\max} \left(\sqrt{K} P \sqrt{K} \right).$$

From the proof of Theorem 2.1 it can be gathered that if $\sigma \in \Delta$ then,

$$\max_{Y \in \mathbb{R}^{n \times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\operatorname{Var}(\hat{\tau}^{CR}|X,Y)} = \left(1 - \frac{1}{n}\right) \lambda_{\max}(P(\sigma)).$$

Therefore, if we wish, we may ensure that we do not stray too far from complete randomization in the worst realization of outcomes by instead solving

$$\begin{aligned} \frac{4}{n^2} \min_{P \in \mathcal{P}} \quad & \lambda_{\max} \left(\sqrt{K} P \sqrt{K} \right) \\ \text{s.t.} \quad & \left(1 - \frac{1}{n}\right) \lambda_{\max}(P) \leq \rho. \end{aligned} \tag{4.2}$$

Since setting $\rho = \infty$ eliminates the constraint, we will only treat (4.2) as it is most general. Setting $\rho = 1$ forces (4.2) to choose complete randomization.

While the problem (4.2) has a convex objective and convex feasible region, we have already observed in Section 2.4 that the problem is NP-hard. When $\rho = \infty$, the feasible region is $\mathcal{P}$, which is a polytope. But what makes (4.2) with $\rho = \infty$ more difficult than the problem encountered in Section 4.1.3 is that, at the same time as being NP-hard, it is not amenable to the branch-and-bound techniques employed by integer optimization software because its optimum generally does not occur at a corner point of the polytope, as we observed in Section 2.4. The polytope $\mathcal{P}$ is known as the equipartition polytope of the complete graph on n vertices [40, 41].

Therefore, we propose only heuristic solutions to the problem. These heuristics are based on semi-definite optimization (SDO), i.e., optimization over the cone S_+^n of $n \times n$ positive semi-definite matrices (see [42] for more information on SDO). In particular, the heuristics run in polynomial time. We use Mosek [43] to solve all SDO problems in our numerical experiments.

The first heuristic is based on a semidefinite outer approximation $\mathcal{P} \subset \left\{P \in S_+^n : \operatorname{diag}(P) = e,\ Pe = 0\right\}$ and is motivated by [44] and [45].

ALGORITHM 4.1. Let $\hat{P}$ be a solution to the SDO

$$\begin{aligned} \min_{\lambda \in \mathbb{R},\, P \in S_+^n} \quad & \lambda \\ \text{s.t.} \quad & \lambda I - \sqrt{K} P \sqrt{K} \in S_+^n \\ & \rho I - \left(1 - \frac{1}{n}\right) P \in S_+^n \\ & \operatorname{diag}(P) = e,\ Pe = 0. \end{aligned}$$

Let $\hat{\sigma}$ be the distribution of $u_i = \operatorname{sign}\left(v_i - \operatorname{median}(v)\right)$ where $v \sim \mathcal{N}(0, \hat{P})$. (This provides a sampling mechanism without needing to fully specify $\hat{\sigma}$).

The second heuristic is based on an inner approximation of $\mathcal{P}$.

ALGORITHM 4.2. Given $u_1, \ldots, u_T \in \mathcal{U}$, let $\hat{\theta}$ be the solution to the SDO

$$\begin{aligned} \min_{\lambda \in \mathbb{R},\ \theta \in \mathbb{R}^T} \quad & \lambda \\ \text{s.t.} \quad & \lambda I - \sum_{t=1}^{T} \theta_t \sqrt{K} u_t u_t^T \sqrt{K} \in S_+^n \\ & \rho I - \left(1 - \frac{1}{n}\right) \sum_{t=1}^{T} \theta_t u_t u_t^T \in S_+^n \\ & \theta \geq 0,\ \sum_{t=1}^{T} \theta_t = 1. \end{aligned}$$

Let $\hat{\sigma}$ be the distribution of $u = \pm u'$ equiprobably where u' is drawn randomly from $\{u_t\}$ according to weights $\hat{\theta}$.

The inputs to Algorithm 4.2 can be generated in two ways. One way is to run Algorithm 4.1 and use the solution to draw u_t (filtering non-unique values up to negation). Another way is to use as inputs the top T solutions to the pure-strategy problem. As this is the method we use in our numerical experiments we describe it explicitly below.

ALGORITHM 4.3. Let $\mathcal{U}_1 = \mathcal{U} \cap \{u_1 = 1\}$. For $t = 1, \ldots, T$ do:

1: Solve $u_t \in \arg\min_{u \in \mathcal{U}_t} u^T K u$.
2: Set $\mathcal{U}_{t+1} = \mathcal{U}_t \cap \left\{u_t^T u \leq n - 4\right\}$.

Run Algorithm 4.2 using $u_1, \ldots, u_T$.

The definition of $\mathcal{U}_1$ simply eliminates the symmetry of negation. Each further refinement in step 2 cuts away the last optimal solution.

5. Algorithms for inference. A priori balance has the potential to significantly reduce estimation variance. One would expect therefore that inferences on the treatment effect can also have higher statistical power. In this section, we will consider $m = 2$ and the sharp null hypothesis

$$H_0 : (\mathrm{TE}_i = 0 \ \forall i = 1, \ldots, n) .$$

Under H_0 all post-treatment responses are exchangeable regardless of treatment given $(Y_{i1} = Y_{i2})$. We can therefore simulate what would happen under another assignment and compare. This is the idea behind Fisher's randomization test, where new simulated assignments are drawn from the same design as used at the onset of the experiment. However, the pure-strategy optimal design when $\mathcal{F}$ is an RKHS generally only randomizes over treatment-permutations of a single partition, which does not provide enough comparison (applying Fisher's randomization test will always yield $p = 1$). Therefore, we develop an alternative test based on the bootstrap [46]:

ALGORITHM 5.1. For a confidence level $0 < 1 - \alpha < 1$:

1: Draw W^0 from the pure-strategy optimal design for the baseline covariates $X_1, \ldots, X_n$, assign subjects, apply treatments, measure outcomes $Y_{iW_i^0}$, and compute $\hat{\tau}$.
2: For $t = 1, \ldots, T$ do:
  2.1: Sample $i_j^t \sim \mathrm{Unif}\{1, \ldots, n\}$ independently for $j = 1, \ldots, n$.
  2.2: Draw W^t from the pure-strategy optimal design for the baseline covariates $X_{i_1^t}, \ldots, X_{i_n^t}$.
  2.3: Compute $\tilde{\tau}^t = \frac{1}{p} \sum_{i:W_i^t=1} Y_{iW_i^0} - \frac{1}{p} \sum_{i:W_i^t=2} Y_{iW_i^0}$.
  (Notice we only use the outcomes we chose to observe in step 1.)
3: The p-value of H_0 is $p = \left(1 + \left|\left\{t \ : \ \left|\tilde{\tau}^t\right| \geq |\hat{\tau}|\right\}\right|\right) / (1 + T)$.
If $p \leq \alpha$, then reject H_0.

Algorithm 5.1 can also be used to answer inferential questions for mixed-strategy designs, letting W^t be drawn from the corresponding mixed-strategy optimal design σ^t in step 2.2. However, the additional randomization of mixed-strategy optimal designs (and of complete randomization, blocking, pairwise matching, and re-randomization for that matter) allows one to use

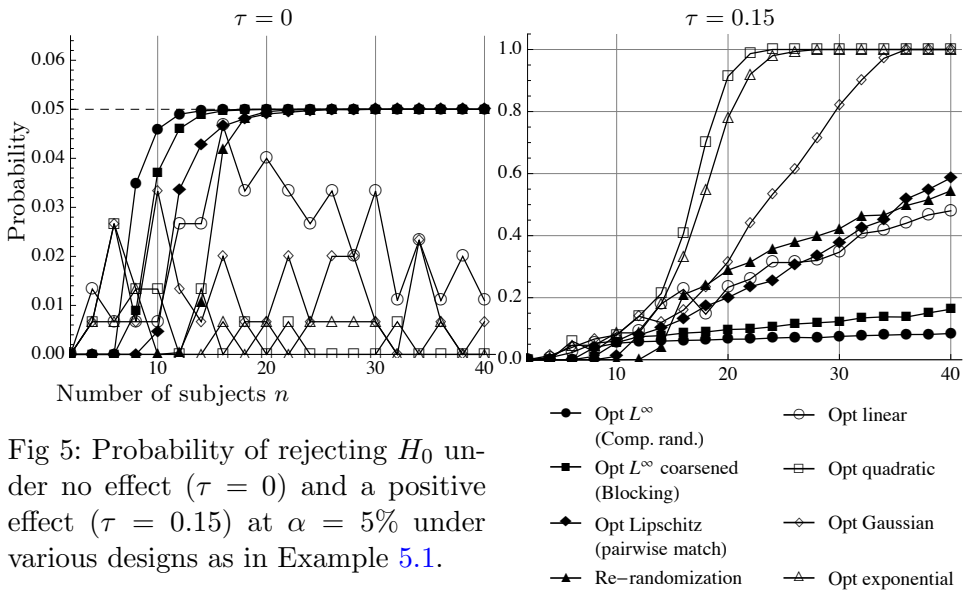

Fig 5: Probability of rejecting H_0 under no effect ($\tau = 0$) and a positive effect ($\tau = 0.15$) at $\alpha = 5\%$ under various designs as in Example 5.1.

the standard randomization and exact permutation tests instead (where new assignments are drawn from the same design as used at the onset of the experiment). As these tests are standard we defer further discussion to supplemental Section 7.3. We next consider an example using Algorithm 5.1.

EXAMPLE 5.1. Consider the setup as in Example 2.2 with $d = 2$, quadratic $\hat{f}$, and $\epsilon_{i1} = \epsilon_{i2} = 0$. For various values of τ, we test H_0 at significance $\alpha = 0.05$ for each of the designs in Example 2.2 (replacing the mixed-strategy optimal designs with corresponding pure-strategy optimal designs) using Algorithm 5.1 for all RKHS-based optimal designs and the standard randomization test for all other designs (see Algorithm 7.2 in supplemental Section 7.3). We plot in Figure 5 the probability of rejecting H_0 as n grows.

When τ is positive, the quadratic and exponential RKHS-based designs detect the difference in treatments almost immediately, the Gaussian a bit later. The linear RKHS-based design parametrically misspecifies the regression function in this particular case but does not do much worse than the other designs nonetheless. Interestingly, as imbalance disappears, Algorithm 5.1 has much lower type I error than the significance $\alpha = 0.05$.

6. Concluding remarks. Designs that provide balance in controlled experiments before treatments are applied and before randomization provide one answer to the criticism that complete randomization may lead to assignments that the experimenter knows will lead to misleading conclusions. In

this paper we unified these designs under the umbrella of a priori balance. We argued that structural information on the dependence of outcomes on baseline covariates was the key to any a priori balance beyond complete randomization and developed a framework of optimal designs based on structure expressed on the conditional expectation function. We have shown how existing a priori balancing designs, including blocking, pairwise matching, and other designs, are optimal for certain structures and how existing imbalance metrics, such as the group-wise Mahalanobis metric of [7], arise from other choices of structure. That this theoretical framework fit so well into existing practice, led us to endeavor to discover what other designs may arise from it. We considered a wide range of designs that follow from structure expressed using RKHS, encompassing both parametric and non-parametric methods. We argued and shown numerically that parametric models (when correctly specified) coupled with optimization lead to estimation variance that converges very fast to the best theoretically possible.

I hope this paper provides a better understanding of the intuitively attractive ambition to always balance experimental groups at the onset of an experiment along with the practical benefits of insights that accompany this understanding and of a wider range of procedures that achieve optimal balance. It has not escaped my notice that this unified perspective on a priori balance suggests a possible rephrasing of Box's maxim: "*balance* what you can, randomize what you cannot."

Acknowledgements. I would like to thank the anonymous reviewers for their helpful comments, my brother Yoav Kallus for interesting conversations and editorial inputs, and Dimitris Bertsimas, a discerning advisor.

7. Supplement.

7.1. *A priori balance in estimating treatment effect on compliers.* In many experimental endeavors involving human subjects the researcher does not fully control the treatment actually administered. Consider two treatments, "treatment" ($k = 1$) and "control" ($k = 2$). Situations where a subject receives a treatment different from their assignment include refusal of surgery, ethical codes that allow subjects assigned to control to demand treatment, or the leakage of information to some control subjects in a teaching intervention. This issue is termed non-compliance. In such situations, W represents initial assignment intent and our estimator $\hat{\tau}$ estimates the effect of the *intent* to treat (ITT). Often a researcher is interested in the compliers' average treatment effect in the sample (CSATE) or population (CPATE), disregarding all non-compliers. Subjects that always demand treatment are known as always-takers, those that always refuse treatment as never-takers, and those that always choose the opposite of their assignment as defiers (this is exhaustive if subjects comply based only on their own assignment). Denote by π_c and Π_c the unknown fraction of compliers in the sample and population, respectively. In the absence of defiers we can observe the identity of never-takers in the treatment group and of always-takers in the control group. We can estimate the fraction of compliers as the complement of those:

$$\hat{\pi}_c = 1 - \frac{2}{n} \sum_{i:W_i=1} \mathrm{NT}_i - \frac{2}{n} \sum_{i:W_i=2} \mathrm{AT}_i$$

where $\mathrm{NT}_i = 1$ if i is a never-taker and $\mathrm{AT}_i = 1$ if i is an always-taker (both 0 for compliers). Under an assignment that blinds the identity of treatment, such as complete randomization, $\hat{\pi}_c$ is conditionally (for π_c) and marginally (for Π_c) unbiased if there are no defiers. Moreover, without defiers,

$$\mathrm{CSATE} = \mathrm{SATE}\,/\pi_c \qquad \mathrm{CPATE} = \mathrm{PATE}\,/\Pi_c$$

since the individual ITT effect for an always- or never-taker is identically 0. It has been often advocated (see [47, 48]) in completely randomized trials to estimate the compliers' average treatment effect by a ratio estimator $\hat{\tau}_c = \hat{\tau}/\hat{\pi}_c$. Such an estimator need not be unbiased but because it is the ratio of two unbiased estimators it has been argued to be approximately unbiased (ibid.). Under a design that blinds the identity of treatments the two estimators remain unbiased and the very same approach can be taken.

We can do even better if we use a priori balance to improve the precision of the compliance fraction estimator. The difference between the sample

compliance fraction and our estimator of it can be seen to be

$$\hat{\pi}_c - \pi_c = \frac{2}{N}\sum_{i:W_i=1}(\mathrm{AT}_i - \mathrm{NT}_i) - \frac{2}{n}\sum_{i:W_i=2}(\mathrm{AT}_i - \mathrm{NT}_i) = \frac{2}{n}\sum_{i=1}^{n} u_i C_i$$

$$\text{where} \quad C_i = \begin{cases} 1 & i \text{ is always-taker} \\ 0 & i \text{ is complier} \\ -1 & i \text{ is never-taker} \end{cases} \quad \text{is } i\text{'s compliance status.}$$

Therefore, matching the means of $f_c(x) = \mathbb{E}\left[C_i | X_i = x\right]$ will eliminate variance in estimating the compliance fraction and get us closer to the true CSATE and CPATE. Moreover, if the two unbiased estimators, $\hat{\tau}$ and $\hat{\pi}_c$, are both more precise, their ratio $\hat{\tau}_c$ is both more precise and less biased. To achieve this through our framework we need only incorporate our belief $\mathcal{F}_c$ about f_c into the larger $\mathcal{F}$ and proceed as before. (See also supplemental Section 7.2 for a discussion about combining spaces.)

7.2. *Generalizations of $\mathcal{F}$.* In this supplemental section we consider more general forms of the space $\mathcal{F}$. For the most part, the theorems presented in the main text will still apply. We deferred this discussion to this supplement to avoid overly cumbersome notation in the main text.

First, we consider the restriction to cones in $\mathcal{F}$. A cone is a set $C \subset \mathcal{F}$ such that $f \in C \implies cf \in C \ \forall c > 0$. We may then further restrict to $f \in C$, $||f|| \leq 1$ in the definitions of $M^2_{\mathrm{p}}(W)$ and $M^2_{\mathrm{m}}(\sigma)$. By symmetry, this is the same as restricting to $C \cup (-C)$. Since it is still the case that $||cf|| = c\,||f||$, Theorems 3.2 and 3.4 still apply. One example of a cone is the cone of monotone functions (either nondecreasing or nonincreasing). In a single dimension and for two treatments, this will result in a pure-strategy optimal design that sorts the data and assigns subjects in an alternating fashion. This is also a feasible assignment for pairwise matching in one dimension. More generally and in higher dimensions, we can consider a directed acyclic graph (DAG) on the nodes $V = \{1, \ldots, n\}$ with edge set $E \subset V^2$ and its associated topological cone $C = \{f : f(X_i) \leq f(X_j)\ \forall (i,j) \in E\}$. Other cones include nonnegative/positive functions and $\pm$-sum-of-squares polynomials.

Second, we consider re-centering the norms. We might have a nominal regression function g that we believe is approximately right, perhaps due to a prior regression analysis or based on models from the literature. In this case, it would make sense to solve the minimax problem against perturbations around this g. Given a norm $||\cdot||'$ on $\mathcal{F}$ we can formally define the magnitude

$$||f|| = \max\left\{\min\left\{||f-g||', ||f+g||'\right\}, 1\right\}. \tag{7.1}$$

We consider both g and $-g$ because it has no effect on the imbalance metrics due to symmetry of the objective while it can only reduce magnitudes. Using this alternate definition of $||\cdot||$ in (7.1), Theorem 3.2 still applies and Theorem 3.4 applies if its conditions apply to the Banach space $\mathcal{F}$ with its usual norm and $\mathbb{E}\,|g(X_1)| < \infty$. In the Bayesian interpretation discussed in Section 2.4.1, this is equivalent to making the prior mean of $f(x)$ be $g(x)$.

Third, we consider combining multiple spaces $\mathcal{F}_1, \ldots, \mathcal{F}_b$. There are two ways. The first way is to combine these via an algebraic sum. The space $\mathcal{F} = \mathcal{F}_1 + \cdots + \mathcal{F}_b = \{\phi_1 + \cdots + \phi_b : \phi_j \in \mathcal{F}_j \ \forall j\}$ endowed with the norm $||f|| = \min_{\phi_j \in \mathcal{F}_j : f = \phi_1 + \cdots + \phi_b} \max_{j=1,\ldots,b} ||\phi_j||_{\mathcal{F}_j}$ is Banach space and as such a valid choice. In particular, the algebraic sum $\mathcal{F}$ can be identified with the quotient of the direct sum $\mathcal{F}' = \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_b$ by its subspace $\{(\phi_1, \ldots, \phi_b) \in \mathcal{F}' : \phi_1 + \cdots + \phi_b = 0\}$. We can decompose the pure-strategy imbalance metric corresponding to this new choice as follows:

$$M_{\mathrm{p}}^2(W) = \max_{k \neq k'} \left(\sum_{j=1}^{b} \sup_{||\phi_j||_{\mathcal{F}_j} \leq 1} B_{kk'}(W, \phi_j) \right)^2 .$$

Theorems 3.2 and 3.4 still apply (in particular the conditions of Theorem 3.4 hold for $\mathcal{F}$ if they hold for each $\mathcal{F}_j$).

The second way is to combine these formally via a union. Consider the space $\mathcal{F} = \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_b = \{f : f \in \mathcal{F}_j \text{ for some } j\}$. This is not a vector space but we can formally define the magnitude $||f|| = \min_{j=1,\ldots,b} ||f||_{\mathcal{F}_j}$. We can then decompose the pure-strategy imbalance metric corresponding to this new choice as follows:

$$M_{\mathrm{p}}^2(W) = \max_{k \neq k'} \max_{j=1,\ldots,b} \sup_{||\phi_j||_{\mathcal{F}_j} \leq 1} B_{kk'}^2(W, \phi_j).$$

Theorem 3.2 still applies and Theorem 3.4 applies if its conditions hold for each Banach space $\mathcal{F}_j$.

We can even take several spaces $\mathcal{F}_1, \ldots, \mathcal{F}_b$, re-center each norm with its own g_j as in (7.1), and then combine them in either of the two ways, defining the combined magnitudes strictly formally. In this way, we can have multiple centers to represent various beliefs about the same or different regression functions f_k. Theorem 3.2 still applies and Theorem 3.4 applies if its conditions hold for each $\mathcal{F}_j$ and $\mathbb{E}\,|g_j(X_1)| < \infty$ for for each j.

7.3. *Inference for mixed-strategy designs.* As noted in Section 5 Algorithm 5.1 can be used to answer inferential questions for mixed-strategy

designs as well, but their additional randomization allows for the standard randomization and exact permutation tests to be used instead. The following is the standard permutation test when applied to a non-completely randomized design, including the mixed-strategy optimal design.

ALGORITHM 7.1. Let σ be given. For a confidence level $0 < 1-\alpha < 1$:

1: Draw W^0 from σ, assign subjects, apply treatments, measure $Y_{iW_i^0}$, and compute $\hat{\tau}$. Let $\mathcal{W}' = \{W \in \mathcal{W} : \sigma(W) > 0\}$.
2: For $W \in \mathcal{W}'$ compute $\tilde{\tau}^W = \frac{1}{p}\sum_{i:W_i=1} Y_{iW_i^0} - \frac{1}{p}\sum_{i:W_i=2} Y_{iW_i^0}$.
3: The p-value of H_0 is $p = \sum_{W\in\mathcal{W}'} \sigma(W)\mathbb{I}\left[\left|\tilde{\tau}^W\right| \geq |\hat{\tau}|\right]$.
If $p \leq \alpha$ then reject H_0.

The above exact test requires that we have a full description of σ and that we iterate over all feasible assignments. This works well for the output of Algorithm 4.2 but can be prohibitive for the output of Algorithm 4.1. The standard randomization test eschews these issues.

ALGORITHM 7.2. Let σ be given. For a confidence level $0 < 1-\alpha < 1$:

1: Draw W^0 from σ, assign subjects, apply treatments, measure $Y_{iW_i^0}$, and compute $\hat{\tau}$.
2: For $t = 1, \ldots, T$ do:
2.1: Draw W^t from σ.
2.2: Compute $\tilde{\tau}^t = \frac{1}{p}\sum_{i:W_i^t=1} Y_{iW_i^0} - \frac{1}{p}\sum_{i:W_i^t=2} Y_{iW_i^0}$.
3: The p-value of H_0 is $p = \left(1 + \left|\left\{t \ : \ \left|\tilde{\tau}^t\right| \geq |\hat{\tau}|\right\}\right|\right) / (1+T)$.
If $p(H_0) \leq \alpha$ then reject H_0.

7.4. *Proofs.*

PROOF OF THEOREM 2.1. Simple arithmetic yields,

$$\hat{\tau} - \text{SATE} = \frac{2}{n}\sum_{i:W_i=1}\left(\frac{Y_{i1}+Y_{i2}}{2}\right) - \frac{2}{n}\sum_{i:W_i=2}\left(\frac{Y_{i1}+Y_{i2}}{2}\right) = \frac{2}{n}\sum_{i=1}^{n} u_i\hat{Y}_i.$$

By conditional unbiasedness, we have

$$\text{Var}\left(\hat{\tau}|X,Y\right) = \mathbb{E}\left[\left(\hat{\tau}-\text{SATE}\right)^2\Big|X,Y\right] = \mathbb{E}\left[\left(\frac{2}{n}\sum_{i=1}^{n} u_i\hat{Y}_i\right)^2\Bigg|X,Y\right].$$

Consider any feasible $\sigma \in \Delta$ and let W be drawn from it. Because shifting Y_1 by one constant and Y_2 by another amounts to shifting $\hat{Y}$ by a constant, which does not change $\hat{\tau}$, by minimizing norms we have that

$$\max_{Y \in \mathbb{R}^{n\times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\|Y_1\|^2 + \|Y_2\|^2} = \max_{Y \in \mathbb{R}^{n\times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\|\overline{Y}\|^2} = \max_{Y \in \mathbb{R}^{n\times 2}} \frac{\operatorname{Var}(\hat{\tau}|X,Y)}{\|\hat{Y}\|^2}$$

$$= \max_{\hat{Y} \in \mathbb{R}^n : \|\hat{Y}\| \le 1} \sum_{W \in \mathcal{W}} \sigma(W) \left(\frac{2}{n} \sum_{i=1}^{n} u_i \hat{Y}_i \right)^2. \tag{7.2}$$

Suppose $\sigma \in \Delta$ minimizes (7.2). For any $\pi \in S_n$ a permutation of $\{1, \ldots, n\}$, define $\sigma_\pi((W_1, \ldots, W_n)) = \sigma((W_{\pi(1)}, \ldots, W_{\pi(n)}))$. Then by the symmetry of $||\cdot||$, σ_π is also optimal. Next note that (7.2) is a maximum over linear forms in σ and is therefore convex. Therefore, $\sigma^*(W) = \frac{1}{n!} \sum_{\pi \in S_n} \sigma_\pi(W)$ is also optimal. By construction we get $\sigma^*((W_1, \ldots, W_n)) = \sigma^*((W_{\pi(1)}, \ldots, W_{\pi(n)}))$ for any $\pi \in S_n$. Hence, $\sigma^*((W_1, \ldots, W_n)) = \sigma^*((1, 2, 1, 2, \ldots, 1, 2))$ is constant for every $W \in \mathcal{W}$, and therefore σ^* is complete randomization. □

Proof of Theorem 2.2. First note that by (2.2), for any i, j, k, k',

$$\sigma\left(\{W_i = W_j,\, W_i \in \{k, k'\},\, W_j \in \{k, k'\}\}\right) = \frac{2}{m}\sigma\left(\{W_i = W_j\}\right),$$
$$\sigma\left(\{W_i \neq W_j,\, W_i \in \{k, k'\},\, W_j \in \{k, k'\}\}\right) = \frac{2}{m}\frac{1}{m-1}\sigma\left(\{W_i \neq W_j\}\right).$$

Therefore, by squaring and interchanging sums, we have

$$M_{\mathrm{m}}^2(\sigma) = \max_{||f|| \le 1} \max_{k \neq k'} \frac{1}{p^2} \sum_{i,j=1}^{n} f(X_i) f(X_j) \sum_{W \in \mathcal{W}} \sigma(W)(w_{ik} - w_{ik'})(w_{jk} - w_{jk'})$$
$$= \max_{||f|| \le 1} \max_{k \neq k'} \frac{2}{pn} \sum_{i,j=1}^{n} P_{ij}(\sigma) f(X_i) f(X_j)$$
$$= \max_{||f|| \le 1} \frac{2}{pn} \sum_{i,j=1}^{n} P_{ij}(\sigma) f(X_i) f(X_j). \qquad \square$$

Proof of Theorem 2.3. Let $\{x_1, \ldots, x_\ell\}$ be the set of values taken by the baseline covariates $X_1, \ldots, X_n$ ($\ell \le n$). Let an assignment W be given. Let $\{i_1, i_1'\}, \ldots, \{i_q, i_q'\}$ denote a maximal perfect exact match across the two groups ($W_{i_j} = 1$, $W_{i_j'} = 2$, $X_{i_j} = X_{i_j'}$, and q maximal) with $\{i_1'', \ldots, i_{q'}''\}$, $\{i_1''', \ldots, i_{q'}'''\}$ being the remaining unmatched subjects ($W_{i_j''} = 1$, $W_{i_{j'}'''} = 2$,

$X_{i''_j} \neq X_{i'''_{j'}}$). For $i = 1, \ldots, \ell$, if there are more x_i's in group 1 set $f'(x_i) = 1$ otherwise set $f'(x_i) = -1$. This f' is feasible ($||f'||_\infty \leq 1$) and hence

$$\max_{||f|| \leq 1} |B(W, f)| \geq \left|B(W, f')\right| = \frac{2}{n} \times q' \times 2 = 2 - \frac{4}{n} q.$$

At the same time, we have

$$\begin{aligned}
\max_{||f|| \leq 1} |B(W, f)| &= \max_{||f|| \leq 1} \left| \sum_{i=1}^{n} u_i f(X_i) \right| \\
&\leq \frac{2}{n} \sum_{j=1}^{q} \max_{||f|| \leq 1} \left| f(X_{i_j}) - f(X_{i'_j}) \right| + \frac{2}{n} \sum_{j=1}^{q'} \max_{||f|| \leq 1} \left| f(X_{i''_j}) - f(X_{i'''_j}) \right| \\
&= 0 + \frac{2}{n} \times q' \times 2 = 2 - \frac{4}{n} q.
\end{aligned}$$

To summarize,

$$\sqrt{M_{\mathrm{p}}^2(W)} = 2 - \frac{4}{n} \begin{pmatrix} \text{number of perfect exact matches} \\ \text{across the experimental groups} \end{pmatrix}. \qquad \square$$

Proof of Theorem 2.4. Let $D_{ij} = \delta(X_i, X_j)$. The pure-strategy optimal design solves the optimization problem

$$\min_{W \in \mathcal{W}} \max_{||f||_{\mathrm{lip}} \leq 1} |B(W, f)| = \frac{2}{n} \min_{\substack{u \in \{-1,1\}^n \\ \sum_{i=1}^n u_i = 0}} \max_{\substack{y \in \mathbb{R}^n \\ y_i - y_j \leq D_{ij}}} u^T y. \tag{7.3}$$

We will show that the set of optimal solutions u to (7.3) is equal to the set of assignments of $+1, -1$ to the pairs in any minimal-weight pairwise match. Since the pure-strategy optimal design randomizes over these, this will show that it is equivalent to pairwise matching, which randomly splits pairs.

Consider any non-bipartite matching $\mu = \{\{i_1, j_1\}, \ldots, \{i_{n/2}, j_{n/2}\}\}$ and any $t \in \{-1, +1\}^{n/2}$. Let $u_{i_l} = t_l$, $u_{j_l} = -t_l$. Enforcing only a subset of the constraints on y, the cost of u in (7.3) is bounded above as follows

$$\max_{y_i - y_j \leq D_{ij}} u^T y = \max_{y_i - y_j \leq D_{ij}} \sum_{l=1}^{n/2} t_l (y_{i_l} - y_{j_l}) \leq \sum_{l=1}^{n/2} D_{i_l j_l},$$

which is the matching cost of μ. Now let instead a feasible solution u to (7.3) be given. Let $S = \{i : u_i = +1\} = \{i_1, \ldots, i_{n/2}\}$ and its complement $S^C = \{i : u_i = -1\} = \{i'_1, \ldots, i'_{n/2}\}$. By linear programming duality we have

$$\max_{y_i - y_j \leq D_{ij}} u^T y = \min_{Fe - F^T e = u,\, F \geq 0} \sum_{i,j=1}^{n} D_{ij} F_{ij} \tag{7.4}$$

since the LHS is bounded ($\leq D_{i_1 i'_1} + \cdots + D_{i_{n/2} i'_{n/2}}$) and feasible ($y_i = 0\ \forall i$). The RHS is an uncapacitated min-cost transportation problem with sources S (with inputs 1) and sinks S^C (with outputs 1). Consider any $j_s \in S$, $j_t \in S^C$ and any path $j_s, j_1, \ldots, j_p, j_t$. By the triangle inequality,

$$D_{j_s j_t} \leq D_{j_s j_1} + D_{j_1 j_2} + \cdots + D_{j_p j_t}.$$

Therefore, it is always preferable to send flow along edges between S and S^C only. Thus, erasing all edges within S or S^C, the problem is seen to be a bipartite matching problem. The min-weight bipartite matching is also a non-bipartite matching and by (7.4) its matching cost is the same as the cost of the given u in the objective of (7.3). □

PROOF OF THEOREM 2.5. The argument is similar to the above. This time the network flow problem has an additional node with zero external flow (neither sink nor source), uncapacitated edges into it from every other node with a unit cost of δ_0, and uncapacitated edges out of it to every other node with a unit cost of δ_0. □

PROOF OF THEOREM 2.6. For the pure-strategy case we have,

$$\begin{aligned}
M_{\mathrm{p}}^2(W) &= \max_{k \neq k'} \max_{||f|| \leq 1} \left(\frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) f(X_i) \right)^2 \\
&= \max_{k \neq k'} \left|\left| \frac{1}{p} \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \mathcal{K}(X_i, \cdot) \right|\right|^2 \\
&= \frac{1}{p^2} \max_{k \neq k'} \left\langle \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \mathcal{K}(X_i, \cdot), \quad \sum_{i=1}^{n} (w_{ik} - w_{ik'}) \mathcal{K}(X_i, \cdot) \right\rangle \\
&= \frac{1}{p^2} \max_{k \neq k'} \sum_{i,j=1}^{n} (w_{ik} - w_{ik'}) K_{ij} (w_{jk} - w_{jk'})
\end{aligned}$$

Now, consider the maximum over f in $M_{\mathrm{m}}^2(P)$. Let f_0 be a feasible solution. Write $f_0 = f + f^{\perp}$ with $f \in S = \operatorname{span}\{\mathcal{K}(X_i, \cdot) \ : \ i = 1, \ldots, n\}$ and $f^{\perp} \in S^{\perp}$, its orthogonal complement. By orthogonality $f^{\perp}(X_i) = \langle \mathcal{K}(X_i, \cdot), f^{\perp} \rangle = 0$ and $||f||^2 = ||f_0||^2 - \left|\left|f^{\perp}\right|\right|^2 \leq 1$ so that f achieves the same objective value as f_0 and remains feasible. Therefore we may restrict to S and assume that $f = \sum_i \beta_i \mathcal{K}(X_i, \cdot)$ such that $\beta^T K \beta \leq 1$.

By positive semi-definiteness of K and P, we get

$$M_{\mathrm{m}}^2(P) = \frac{2}{np} \sup_{\beta^T K \beta \leq 1} \sum_{i,j=1}^{n} P_{ij} (K\beta)_i (K\beta)_j = \frac{2}{np} \sup_{\beta^T K \beta \leq 1} \beta^T K P K \beta$$

$$= \frac{2}{np} \sup_{\gamma^T\gamma \le 1} \gamma^T \sqrt{K} P \sqrt{K} \gamma = \frac{2}{np} \lambda_{\max}\left(\sqrt{K} P \sqrt{K}\right). \qquad \square$$

PROOF OF THEOREM 3.1. By blinding of treatments (2.2), each W_i by itself (but *not* the vector W) is statistically independent of X, Y so that

$$\mathbb{E}\left[w_{ik} Y_{ik} \middle| X, Y\right] = \mathbb{E}\left[w_{ik}\right] Y_{ik} = \frac{1}{m} Y_{ik}, \text{ and therefore}$$
$$\mathbb{E}\left[\hat{\tau}_{kk'} \middle| X, Y\right] = \frac{1}{p}\sum_{i=1}^{n} \frac{1}{m} Y_{ik} - \frac{1}{p}\sum_{i=1}^{n} \frac{1}{m} Y_{ik'} = \mathrm{SATE}_{kk'}.$$

Note that we can rewrite $E_{kk'}$ as

$$E_{kk'} = \frac{1}{m}\sum_{l \ne k} \Xi_{kl} - \frac{1}{m}\sum_{l \ne k'} \Xi_{k'l} \quad \text{where} \quad \Xi_{kl} := \frac{1}{p}\sum_{i=1}^{n} \left(w_{ik} - w_{il}\right) \epsilon_{ik}.$$

Using the notation $A_{kk'} = \frac{1}{p}\sum_{i:W_i=k'} Y_{ik}$, $C_{kl} = B_{kl}\left(f_k\right) + \Xi_{kl}$, we have

$$\begin{aligned}
\hat{\tau}_{kk'} - \mathrm{SATE}_{kk'} &= A_{kk} - A_{k'k'} - \frac{1}{m}\sum_{l=1}^{m} A_{kl} + \frac{1}{m}\sum_{l=1}^{m} A_{k'l} \\
&= \frac{m-1}{m} A_{kk} - \frac{1}{m} A_{kk'} + \frac{1}{m} A_{k'k} - \frac{m-1}{m} A_{k'k'} \\
&\quad - \frac{1}{m}\sum_{l \ne k,k'} \left(A_{kk} - C_{kl}\right) + \frac{1}{m}\sum_{l \ne k,k'} \left(A_{k'k'} - C_{k'l}\right) = D_{kk'} + E_{kk'}.
\end{aligned}$$

Let i, j be equal or unequal, k, k', l, l' equal or unequal. Then,

$$\begin{aligned}
\mathrm{Cov}(w_{il} f_k(X_i),\, w_{jl'} \epsilon_{jk'}) &= \mathbb{E}\left[w_{il} w_{jl'} f_k(X_i) \mathbb{E}\left[\epsilon_{jk'} | X, Z\right]\right] \\
&\quad - \mathbb{E}\left[w_{il} f_k(X_i)\right] \mathbb{E}\left[w_{jl'} \mathbb{E}\left[\epsilon_{jk'} | X, Z\right]\right] = 0 - 0 = 0, \\
\mathrm{Cov}((w_{ik} - w_{il}) f_k(X_i),\, f_{k'}(X_j)) &= \mathbb{E}\left[w_{ik} - w_{il}\right] \mathrm{Cov}\left(f_k(X_i), f_{k'}(X_j)\right) = 0, \\
\mathrm{Cov}((w_{ik} - w_{il}) \epsilon_{ik},\, f_{k'}(X_j)) &= \mathbb{E}\left[w_{ik} - w_{il}\right] \mathrm{Cov}\left(\epsilon_{ik}, f_{k'}(X_j)\right) = 0,
\end{aligned}$$

where the latter two equalities are due to the independence of W_i due to blinding treatments. This proves uncorrelateness. The rest follows from an application of the law of total variance and rearranging terms. $\square$

PROOF OF THEOREM 3.2. Define

$$Z(f, g) = \mathbb{E}\left[\left(\frac{1}{p}\sum_{i=1}^{n} \left(w_{i1} - w_{i2}\right) f(X_i)\right)\left(\frac{1}{p}\sum_{i=1}^{n} \left(w_{i1} - w_{i2}\right) g(X_i)\right)\right].$$

By construction, $Z(f,f) \leq ||f||^2 \mathbb{E}\left[M_{\text{opt}}^2\right]$. By condition (2.2),

$$\begin{aligned}
&\text{Var}\left(B_{kl}(f)\right) = Z(f,f) \quad \text{for } l \neq k,\\
&\text{Cov}\left(B_{kl}(f), B_{kl'}(f)\right) = \frac{1}{2}Z(f,f) \quad \text{for } k,l,l' \text{ distinct},\\
&\text{Cov}\left(B_{kl}(f), B_{k'l'}(g)\right) = \begin{cases} \frac{1}{2}Z(f,g) & \text{for } l = l' \notin \{k,k'\},\\ -\frac{1}{2}Z(f,g) & \text{for } l = k',\ l' \neq k,\\ -\frac{1}{2}Z(f,g) & \text{for } l \neq k',\ l' = k,\\ -Z(f,g) & \text{for } l = k',\ l' = k,\\ 0 & \text{for } k,k',l,l' \text{ distinct}. \end{cases}
\end{aligned}$$

It follows that

$$\begin{aligned}
\text{Var}\left(D_{kk'}\right) &= \frac{1}{m^2}\left(\frac{m^2-m}{2}Z(f_k,f_k) + \frac{m^2-m}{2}Z(f_{k'},f_{k'})\right)\\
&\quad + \frac{m+2}{m^2}Z(f_k,f_{k'})\\
&= \frac{1}{m^2}\left(\frac{m^2}{2} - m - 1\right)\left(Z(f_k,f_k) + Z(f_{k'},f_{k'})\right)\\
&\quad + \frac{1}{m^2}\left(\frac{m+2}{2}\right)Z(f_k+f_{k'},f_k+f_{k'})\\
&\leq \frac{1}{m^2}\left(\frac{m^2}{2} - m - 1\right)\left(\mathbb{E}\left[M_{\text{opt}}^2\right]||f_k||^2 + \mathbb{E}\left[M_{\text{opt}}^2\right]||f_{k'}||^2\right)\\
&\quad + \frac{1}{m^2}\left(\frac{m+2}{2}\right)\mathbb{E}\left[M_{\text{opt}}^2\right]||f_k+f_{k'}||^2\\
&\leq \frac{\left(||f_k|| + ||f_{k'}||\right)^2}{2}\left(1 - \frac{1}{m}\right)\mathbb{E}\left[M_{\text{opt}}^2\right]
\end{aligned}$$

since $||f+g||^2 \leq \left(||f|| + ||g||\right)^2$ and $\left(||f||^2 + ||g||^2\right) \leq \left(||f|| + ||g||\right)^2$. □

Proof of Theorem 3.3. Fix f and g. Using $\left(\sum_{i=1}^b z_i\right)^2 \leq b\sum_{i=1}^b z_i^2$,

$$\begin{aligned}
Z(f,f) &= \mathbb{E}\left(\frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)(f-g)(X_i) + \frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)g(X_i)\right)^2\\
&\leq 2\mathbb{E}\left(\frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)(f-g)(X_i)\right)^2 + 2\mathbb{E}\left(\frac{1}{p}\sum_{i=1}^n \left(w_{i1} - w_{i2}\right)g(X_i)\right)^2\\
&\leq \frac{2}{p^2} \times p \times p \times \frac{2}{m} \times \mathbb{E}((f-g)(X_1))^2 + 2Z(g,g) = \frac{4}{m}||f-g||_2^2 + 2Z(g,g)
\end{aligned}$$

The rest is as in the proof of Theorem 3.2, choosing $g \in \mathcal{F}$. □

PROOF OF THEOREM 3.4. Fix the assignment $W_i' = (i \mod p) + 1$ and let $\xi_i^{(k,k')} : f \mapsto \left(f\left(X_{m(i-1)+k}\right) - f\left(X_{m(i-1)+k'}\right)\right)$. Then, since $\xi_i^{(k,k')}$ is in the continuous dual space $\mathcal{F}^*$, we can write $M_{\mathrm{p}}^2(W') = \max_{k \neq k} T_n^{(k,k')}$ where

$$T_n^{(k,k')} = \sup_{||f|| \leq 1} \left(\frac{1}{p} \sum_{i=1}^{p} \xi_i^{(k,k')}(f) \right)^2 = \left|\left| \frac{1}{p} \sum_{i=1}^{p} \xi_i^{(k,k')} \right|\right|_{\mathcal{F}^*}^2.$$

Note $\xi_i^{(k,k')}$ are independent and identically distributed with expectation 0 (i.e., Bochner integral). B-convexity of $\mathcal{F}$ implies the B-convexity of $\mathcal{F}^*$. By B-convexity and the main result of [49] (or by [50] for the Hilbert case),

$$T_n^{(k,k')} \to 0 \text{ almost surely as } n \to \infty.$$

As there are only finitely many k, k', we have $M_{\mathrm{p}}^2(W') \to 0$ almost surely. By construction, $M_{\text{m-opt}}^2 \leq M_{\text{p-opt}}^2 \leq M_{\mathrm{p}}^2(W')$. Hence, the distance between $\hat{\tau}_{kk'}$ and $\mathrm{SATE}_{kk'} + E_{kk'}$ is $|D_{kk'}| \leq \left(1 - \frac{1}{m}\right)(||f_k|| + ||f_{k'}||)\sqrt{M_{\text{opt}}^2} \to 0$ almost surely. Therefore, as $\mathrm{SATE}_{kk'} + E_{kk'}$ is strongly consistent, so is $\hat{\tau}_{kk'}$. □

References.

[1] Student. Comparison between balanced and random arrangements of field plots. *Biometrika*, pages 363–378, 1938.

[2] James J Heckman. Econometric causality. *International Statistical Review*, 76(1):1–27, 2008.

[3] Donald B Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484), 2008.

[4] Richard McHugh and John Matts. Post-stratification in the randomized clinical trial. *Biometrics*, pages 217–225, 1983.

[5] Ronald A Fisher. *The Design of Experiments*. Oliver and Boyd, 1949.

[6] Robert Greevy, Bo Lu, Jeffrey H Silber, and Paul Rosenbaum. Optimal multivariate matching before randomization. *Biostat.*, 5(2):263–275, 2004.

[7] Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. *Ann. Statist.*, 40(2):1263–1282, 2012.

[8] Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

[9] Stuart J Pocock and Richard Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115, 1975.

[10] Adam Kapelner and Abba Krieger. Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 2014.

[11] Shein-Chung Chow, Mark Chang, et al. Adaptive design methods in clinical trials-a review. *Orphanet J Rare Dis*, 3(11), 2008.

[12] Xing Sam Gu and Paul R Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

[13] Dimitris Bertsimas, Mac Johnson, and Nathan Kallus. The power of optimization over randomization in designing experiments involving small samples. Submitted for publication.

[14] Donald B Rubin. Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.*, 81(396):961–962, 1986.
[15] Paul R Rosenbaum. Interference between units in randomized experiments. *J. Amer. Statist. Assoc.*, 102(477), 2007.
[16] David R Cox. *Planning of Experiments*. Wiley, 1958.
[17] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.
[18] Halsey L Royden. *Real Analysis*, volume 3. Prentice Hall, 1988.
[19] John K Hunter and Bruno Nachtergaele. *Applied Analysis*. World Scientific, 2001.
[20] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, 2004.
[21] Ingo Steinwart, Don Hush, and Clint Scovel. An explicit description of the reproducing kernel hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, 52(10):4635–4643, 2006.
[22] Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, Malte Rasch, and Er Smola. A kernel method for the two sample problem. In *Advances in Neural Information Processing Systems 19*, 2007.
[23] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res*, 13:723–773, 2012.
[24] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statis.*, 41(5):2263–2291, 2013.
[25] Michael R Garey and David S Johnson. *Computers and Intractability*, volume 174. Freeman, 1979.
[26] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
[27] George S Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41(2):495–502, 1970.
[28] Carl E Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
[29] Joseph B Kadane and Teddy Seidenfeld. Randomization in a bayesian perspective. *Journal of statistical planning and inference*, 25(3):329–345, 1990.
[30] Leonard J Savage. The foundations of statistics reconsidered. In *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 575–586. University of California Press, 1961.
[31] Stephan Mertens. A physicist's approach to number partitioning. *Theoret. Comput. Sci.*, 265(1):79–108, 2001.
[32] Christian Borgs, Jennifer Chayes, Stephan Mertens, and Chandra Nair. Proof of the local REM conjecture for number partitioning. I: Constant energy scales. *Random Structures & Algorithms*, 34(2):217–240, 2009.
[33] Christian Borgs, Jennifer Chayes, Stephan Mertens, and Chandra Nair. Proof of the local REM conjecture for number partitioning. II. growing energy scales. *Random Structures & Algorithms*, 34(2):241–284, 2009.
[34] Samuel Kotz, NL Johnson, and DW Boyd. Series representations of distributions of quadratic forms in normal variables. I. central case. *Ann. Math. Statist.*, 38(3):823–837, 1967.
[35] Mohammad Ahsanullah, Valery B Nevzorov, and Mohammad Shakil. *An Introduction to Order Statistics*. Atlantis Press, 2013.
[36] Gurobi Optimization Inc. Gurobi optimizer reference manual. http://www.gurobi.

com, 2013.
[37] Miguel Sousa Lobo, Lieven Vandenberghe, Stephen Boyd, and Hervé Lebret. Applications of second-order cone programming. *Linear Algebra Appl.*, 284(1):193–228, 1998.
[38] Jack Edmonds. Paths, trees, and flowers. *Canad. J. Math.*, 17(3):449–467, 1965.
[39] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific, Belmont, 1997.
[40] Michele Conforti, MR Rao, and Antonio Sassano. The equipartition polytope. i: formulations, dimension and basic facets. *Mathematical Programming*, 49(1):49–70, 1990.
[41] Michele Conforti, MR Rao, and Antonio Sassano. The equipartition polytope. ii: valid inequalities and facets. *Mathematical Programming*, 49(1-3):71–90, 1990.
[42] Stephen Poythress Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
[43] The MOSEK optimization software. http://www.mosek.com/.
[44] Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.
[45] Dimitris Bertsimas and Yinyu Ye. Semidefinite relaxations, multivariate normal distributions, and order statistics. In *Handbook of Combinatorial Optimization*, pages 1473–1491. Springer, 1999.
[46] Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapmen and Hall, 1993.
[47] Guido W Imbens and Donald B Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997.
[48] Roderick J Little and Linda HY Yau. Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods*, 3(2):147, 1998.
[49] Anatole Beck. A convexity condition in Banach spaces and the strong law of large numbers. *Proc. Amer. Math. Soc.*, 13(2):329–334, 1962.
[50] Ying-Xia Chen and Wei-Jun Zhu. Note on the strong law of large numbers in a Hilbert space. *Gen. Math.*, 19(3):11–18, 2011.

Nathan Kallus
77 Massachusetts Ave, E40-149
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
E-mail: kallus@mit.edu
URL: http://www.nathankallus.com