Multivariate tie-breaker designs

Tim P. Morrison[1] **and Art B. Owen**[1]

[1]*Stanford University,*
e-mail: tim@stanford.edu; owen@stanford.edu

Abstract: In a tie-breaker design (TBD), subjects with high values of a running variable are given some (usually desirable) treatment, subjects with low values are not, and subjects in the middle are randomized. TBDs are intermediate between regression discontinuity designs (RDDs) and randomized controlled trials (RCTs). TBDs allow a tradeoff between the resource allocation efficiency of an RDD and the statistical efficiency of an RCT. We study a model where the expected response is one multivariate regression for treated subjects and another for control subjects. We propose a prospective D-optimality, analogous to Bayesian optimal design, to understand design tradeoffs without reference to a specific data set. For given covariates, we show how to use convex optimization to choose treatment probabilities that optimize this criterion. We can incorporate a variety of constraints motivated by economic and ethical considerations. In our model, D-optimality for the treatment effect coincides with D-optimality for the whole regression, and, without constraints, an RCT is globally optimal. We show that a monotonicity constraint favoring more deserving subjects induces sparsity in the number of distinct treatment probabilities. We apply the convex optimization solution to a semi-synthetic example involving triage data from the MIMIC-IV-ED database.

1. Introduction

There are many important settings where a costly or scarce intervention can be made for some but not all subjects. Examples include giving a scholarship to some students (Angrist, Hudson and Pallais, 2014), intervening to prevent child abuse in some families (Krantz, 2022), programs to counter juvenile delinquency (Lipsey, Cordray and Berger, 1981) and sending some university students to remedial English classes (Aiken et al., 1998). There are also lower-stakes settings where a company might be able to offer a perk such as a free service upgrade, to some but not all of its customers. In settings such as these, there is usually a priority ordering of subjects, perhaps based on how deserving they are or on how much they or the investigator might gain from the intervention. We can represent that priority order in terms of a real-valued running variable x_i for subject i.

To maximize the immediate short-term value from the limited intervention, one can assign it only to subjects with $x_i \geqslant t$ for some threshold t. The difficulty with this "greedy" solution is that such an allocation makes it difficult to estimate the causal effect of the treatment. It is possible to use a regression discontinuity design (RDD) in this case, comparing subjects with x_i somewhat larger than t to those with x_i somewhat smaller than t. For details on the RDD, see the comprehensive recent survey by Cattaneo and Titiunik (2022). The RDD can give a consistent nonparametric estimate of the treatment effect at $x = t$ (Hahn, Todd and der Klaauw, 2001), but not at other values of x. The treatment effect can be studied at other values of x under an assumed regression model, but then the variance of the regression coefficients can be large due to the strong dependence between the running variable and the treatment (Gelman and Imbens, 2019).

The RDD is commonly used to analyze observational data for which the investigator has no control over the treatment cutoff. In the settings we consider, the investigators assign the treatment and can therefore employ some randomization. The motivating context makes it costly or even ethically difficult to use a randomized controlled trial (RCT) where the treatment is assigned completely at random without regard to x_i. The tie-breaker design (TBD) is a compromise between the RCT and RDD. It is a triage where subjects with large values of x_i get the treatment, those with small values get the control level and those in between are randomized to either treatment or control.

While the tie-breaker design has been known since Campbell (1969) it has not been subjected to much analysis. Most of the theory for the TBD has been for the setting with a scalar x and models that are linear in x and the treatment (Owen and Varian, 2020; Li and Owen, 2022). In this paper we study the TBD for a vector-valued predictor. Our first motivation is that many use cases for TBDs will include multiple covariates. Second, although multivariate nonparametric regression models are out of the scope of this paper, we believe that TBD regressions are a useful first step in that direction. Third, Gelman, Hill and Vehtari (2020) counsel against RDDs that do not adjust for pre-treatment variables. The multivariate regression setting supports such adjustments.

In a tie-breaker design we have a vector of covariates for subject i and must choose a two-level treatment variable. We then face an atypical experimental design problem where some of the predictors are fixed and not settable by us while one of them is subject to randomization. This is known as the 'marginally restricted design' problem after Cook and Thibodeau (1980) who studied *D*-optimality in such a setting. We consider two such settings. In one setting, the investigator already has the covariate values. The second setting is one step removed, where we consider random covariates that an investigator might later have.

The paper is organized as follows. Section 2 outlines our notation and the regression model. Section 3 introduces our notions of efficiency and D-optimality in the multivariate regression model. Theorem 1 shows that D-optimality for the treatment effect parameters is equivalent to D-optimality for the full regression. To study efficiency for future subjects, we use a prospective D-optimality cri-

terion, adapted from Bayesian optimal design, that maximizes expected future information. Theorem 2 then shows that the RCT is prospectively D-optimal. We also discuss an example involving Gaussian covariates and a symmetric design, which provides relevant intuition. Section 4 finds an expression for the expected short-term gain, with particular focus on this Gaussian case. When the running variable is linear in the covariates, the best linear combination for statistical efficiency is the "last" eigenvector of the covariance matrix of covariates while the best linear combination for short-term gain is the true treatment effect. Section 5 presents a design strategy based on convex optimization to choose treatment probabilities for given covariates and compares the effects of various practical constraints. In particular, we show that a monotonicity constraint in the treatment probabilities yields solutions with a few distinct treatment probability levels. This is consistent with some past results in constrained experimental design (Cook and Fedorov, 1995), but both our proof and the theorem particulars are distinct. Section 6 illustrates this procedure on a hospital data set from MIMIC-IV-ED about which emergency room patients should receive intensive care. Section 7 has a brief discussion of some additional context for our results.

1.1. Tie-breaker designs

The earliest tie-breaker design of which we are aware (Campbell, 1969) had a discrete running variable, and literally broke a tie by randomizing the treatment assignments for those subjects with $x = t$. For an imperfectly measured running variable, we might consider $|x - t| < \Delta$ to not be meaningfully large for some $\Delta > 0$, so randomizing in that window is like breaking ties. This is the approach from Boruch (1975).

When we randomize for x with $|x - t| < \Delta$, then setting $\Delta = 0$ provides an RDD while $\Delta \to \infty$ provides an RCT. The TBD transitions smoothly between these extremes. Many of the early comparisons among these methods use the two-line regression model

$$Y_i = \beta_0 + \beta_1 X_i + \gamma_0 Z_i + \gamma_1 Z_i X_i + \varepsilon_i. \tag{1}$$

Here, Y_i is the response for subject i, $Z_i \in \{-1, 1\}$ is a binary treatment variable, and ε_i is an IID error term. Goldberger (1972) finds for $X_i \sim \mathcal{N}(0, 1)$ that the RDD has about 2.75 times the variance of an RCT. Jacob et al. (2012) consider polynomials in X up to third degree, with or without interactions with Z for Gaussian and for uniform random variables. Neither paper considers tie-breakers. Cappelleri, Darlington and Trochim (1994) study the two-line model above (absent the X-Z interaction) and compare the RDD, RCT and three intermediate TBD designs. They study the sample size required to reject $\gamma = 0$ with sufficient power for those designs and three different sizes of γ. Larger Δ brings greater power.

Owen and Varian (2020) find an expression for the variance of the parameters in the two-line model for x with a symmetric uniform distribution about a

threshold $t = 0$, and also for a Gaussian case. The randomization window is $|x| \leqslant \Delta$, with a larger Δ yielding smaller variance but a lesser short-term gain. Their designs have only three levels (0, 50 and 100 percent) for the treatment probability given x. They show that there is no advantage to using other levels or a sliding scale when x has a symmetric distribution and 50% of the subjects will get the treatment.

Li and Owen (2023) study the two-line regression for a general distribution of x and an arbitrary proportion of treated subjects. They constrain the global treatment probability and a measure of short-term gain. They find that there exists a D-optimal design under these constraints with a treatment probability that is constant within at most four intervals of x values. Moreover, with the addition of a monotonicity constraint, there exists an optimal solution with just two levels corresponding to $x < t'$ and $x > t'$ for some t'.

Kluger and Owen (2023) consider tie-breaker design for nonparametric regression for real-valued treatments x. A tie-breaker design can support causal inference about the treatment effect at points within the randomization window, not just at the threshold $x = t$. At the threshold, the widely use kernel regression methods for RDD become much more efficient if one has sampled from a tie-breaker, because they can use interpolation to $x = t$ instead of the extrapolation from the left and right sides that an RDD requires.

1.2. Marginally constrained experimental design

The tie-breaker setup is a special case of marginally constrained experimentation of Cook and Thibodeau (1980). Some more general constraints are considered in Lopez-Fidalgo and Garcet-Rodriguez (2004). Some of our findings are special cases of more general results in Nachtsheim (1989). Heavlin and Finnegan (1998) consider a sequential design setting in which the columns of a design matrix correspond to steps in semi-conductor fabrication, with each step constrained by the prior ones. The marginally constrained literature generally considers choosing n_i values of the settable variables for level i of the fixed variables. In our setting, we cannot assume that $n_i > 1$.

The TBD setting often has a monotonicity constraint on treatment probabilities that we have not seen in the marginally constrained design literature. Under that constraint a more "deserving" subject should never have a lower treatment probability than a less deserving subject has. This and other constraints in the TBD will often lead to optimal designs with a small number of different treatment levels.

1.3. Sequential experimentation

We anticipate that the tie-breaker design could be used in sequential experimentation. In our motivating applications, the response is measured long enough after the treatment (e.g., six years in some educational settings) that bandit

methods (Slivkins, 2019) are not appropriate. There is related work on adaptive experimental design. See for instance Metelkina and Pronzato (2017). The problem we focus on is designing the first experiment that one might use, and a full sequential analysis is not in the scope of this paper.

2. Setup

Given n subjects with d covariates each, we let $X \in \mathbb{R}^{n \times d}$ be the design matrix and $X_i \in \mathbb{R}^d$ be the variables for subject i. We write $\tilde{X} = \begin{bmatrix} 1 & X \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}$ to denote the matrix with an intercept out front, and $\tilde{X}_i = \begin{bmatrix} 1 & X_i^\top \end{bmatrix}^\top \in \mathbb{R}^{d+1}$ to denote its i'th row. For ease of notation, we zero-index $\tilde{X}$ so that $\tilde{X}_{i0} = 1$ and $\tilde{X}_{ij} = X_{ij}$, for $j = 1, \dots, d$.

We are interested in the effect of some treatment $Z_i \in \{-1, 1\}$ on a future response $Y_i \in \mathbb{R}$ for subject i, where $Z_i = 1$ corresponds to treatment and $Z_i = -1$ to control. The design problem is to choose probabilities $p_i \in [0, 1]$ and then take $\mathbb{P}(Z_i = 1) = p_i$. This differs from the common experimental design framework in which the covariates X_i can also be chosen. In Section 5 we will show how to get optimal p_i by convex optimization.

To get more general insights into the design problem, in Section 3 we also consider a random data framework. The predictors are to be sampled with $X_i \overset{\text{iid}}{\sim} P_X$. This allows us to relate design findings to the properties of P_X rather than to a specific matrix X. After X_i are observed, Z_i will be set randomly and then Y_i observed. The analyst is unable to alter P_X at all but can choose any function $p(X) = \mathbb{P}(Z = 1 \mid X)$ that satisfies the imposed constraints.

We work with the following linear model:

$$Y_i = \tilde{X}_i^\top \tilde{\beta} + Z_i \tilde{X}_i^\top \tilde{\gamma} + \varepsilon_i \tag{2}$$

for $\tilde{\beta}, \tilde{\gamma} \in \mathbb{R}^{d+1}$, where ε_i are IID noise terms with mean zero and variance $\sigma^2 > 0$. We use the same notational convention of writing $\tilde{\beta} = \begin{bmatrix} \beta_0 & \beta^\top \end{bmatrix}^\top$ and $\tilde{\gamma} = \begin{bmatrix} \gamma_0 & \gamma^\top \end{bmatrix}^\top$ for $\beta, \gamma \in \mathbb{R}^d$ to separate out the intercept term. We consider $\tilde{\gamma}$ to be the parameter of greatest interest because it captures the treatment effect of Z.

Equation (2) generalizes the two-line model (1) studied by Li and Owen (2023) and Owen and Varian (2020). The latter authors describe some computational methods for the model (2) but most of their theory is for model (1).

Though a strong parametric assumption, this multi-line regression is a helpful and practical model to inform treatment assignment at the design stage. Because our scenario of interest could include regions of the covariate space with $p(x) = 0$, we do not have overlap and cannot rely on classical semiparametric methods for causal inference. Nonparametric generalizations of (2) such as spline models may prove more flexible in highly nonlinear settings, but we do not explore them here.

In Section 3, we also consider multivariate tie-breaker designs (TBDs), which

we define as follows: the treatment Z_i is assigned via

$$
\mathbb{P}(Z_i = 1 \,|\, X_i) = \begin{cases} 1, & X_i^\top \eta \geqslant u \\ p, & X_i^\top \eta \in (\ell, u) \\ 0, & X_i^\top \eta \leqslant \ell \end{cases} \tag{3}
$$

for parameters $u < \ell$, $p \in (0,1)$ and $\eta \in \mathbb{R}^d$. That is, we assign treatment to subject i whenever $X_i^\top \eta$ is at or above some upper cutoff u, do not assign treatment whenever it is at or below some lower cutoff ℓ, and randomize at some fixed probability p in the middle. For the case $u = \ell$, we take $\mathbb{P}(Z_i = 1 \,|\, X_i) = \mathbb{1}\{X_i^\top \eta \geqslant u\}$ which has a mild asymmetry in offering the treatment to those subjects, if any, that have $X_i^\top \eta = \ell = u$.

In equation (3) we ignore the intercept and just consider X_i instead of $\tilde{X}_i$ since the intercept would merely shift everything by a constant. Here, we use $\eta \in \mathbb{R}^d$ instead of γ from (2) to reflect that the vector we treat on need not be the same as the true γ, which is unknown.

In practice, η will encode whatever constraints on randomization a practitioner may encounter. While linear constraints do not encompass all possibilities, they are simple and flexible enough to account for a great many that one could feasibly seek to impose; for example, constraints on a heart rate below some cutoff or a total SAT score that is sufficiently high can both be covered by (3). One could also set ℓ and u to be known or estimated quantiles of a covariate or a linear combination of them if, for example, one wants deterministic treatment assignment for the top 5% of candidates. More generally, we show in Section 5 how to accommodate any convex constraints on treatment assignment.

The assignment (3) greatly generalizes the one in Owen and Varian (2020) which had $d = 1$, $\ell = -u$, $p = 1/2$, and P_X either $\mathcal{U}(-1, 1)$ or $\mathcal{N}(0, 1)$. In analogy to the one-dimensional case, we refer to the case with $u = \ell$ as an RDD. We refer to any choice of (ℓ, u, η) for which $\mathbb{P}(X^\top \eta \in (\ell, u)) = 1$ as an RCT.

3. Efficiency and D-optimality

We assume that the covariates $X_i \in \mathbb{R}^d$ have yet to be observed and are drawn from some distribution P_X with a finite and invertible covariance matrix Σ. The aim is to devise a treatment assignment scheme $p(X)$ with Z_i assigned independently via $\mathbb{P}(Z_i = 1 \mid X_i) = p(X_i)$ that is optimal in some sense. Random allocation of Z has advantages of fairness and supports some randomization-based inference.

Section 3.1 briefly outlines efficiency and D-optimality in the fixed Z setting to introduce relevant formulas. In Section 3.2, we then outline a notion of prospective D-optimality used when X is random and we must choose a distribution of Z given X. We model our approach on the Bayesian optimal design literature (Chaloner and Verdinelli, 1995), in which the model parameters are treated as random, which we describe in more detail in that section.

The treatment of X as random permits a theoretical discussion about what is expected to happen depending on the distributional properties of the unseen X data. However, the case in which the covariates are actually fixed is also covered by the procedure in Section 3.2, since it corresponds to a point mass distribution on X. We return to this case in Section 5, in which we show how to use convex optimization to obtain prospectively D-optimal $p(X_i)$.

3.1. *D-optimality for nonrandom X and Z*

In this brief section, we outline our notions of efficiency and D-optimality in the setting in which X is fixed and Z are assigned non-randomly. While not the focus of this paper, it will allow us to motivate various definitions and derive a property of D-optimality in the multivariate model that will be useful going forward.

We begin by conditioning on X and Z. Let $D \in \mathbb{R}^{n \times n}$ be the diagonal matrix whose diagonal entries are $D_{ii} = Z_i$. We can write the linear model (2) in matrix form as $Y = U\delta + \varepsilon$, where $U = \begin{bmatrix} \tilde{X} & D\tilde{X} \end{bmatrix}$ and $\delta = \begin{bmatrix} \tilde{\beta}^\top & \tilde{\gamma}^\top \end{bmatrix}^\top$.

In the general model (2), conditionally on X and Z we have

$$\mathrm{Var}(\hat{\delta} \,|\, X, Z) = \begin{bmatrix} \mathrm{Var}(\hat{\beta} \,|\, X, Z) & \mathrm{Cov}(\hat{\beta}, \hat{\gamma} \,|\, X, Z) \\ \mathrm{Cov}(\hat{\gamma}, \hat{\beta} \,|\, X, Z) & \mathrm{Var}(\hat{\gamma} \,|\, X, Z) \end{bmatrix}$$

$$\equiv \sigma^2 \begin{bmatrix} ((U^\top U)^{-1})_{11} & ((U^\top U)^{-1})_{12} \\ ((U^\top U)^{-1})_{21} & ((U^\top U)^{-1})_{22,} \end{bmatrix}.$$

where we assume here that U is full rank. Because σ^2 is merely a multiplicative factor independent of all relevant parameters, it is no loss of generality to take $\sigma^2 = 1$ going forward for simplicity. The treatment effect vector $\tilde{\gamma}$ is our parameter of primary interest, so we want to minimize a measure of the magnitude of $((U^\top U)^{-1})_{22}$. When X is fixed, a standard choice would be to assign $\{Z_1, \ldots, Z_n\}$ to minimize the D-optimality criterion of

$$\det\big(((U^\top U)^{-1})_{22}\big) = \prod_{j=1}^{d+1} \lambda_j(\mathrm{Var}(\hat{\gamma} \,|\, Z)),$$

where $\lambda_j(\cdot)$ is the jth eigenvalue of its argument. D-optimality is the most studied design choice among many alternatives (Atkinson, Donev and Tobias, 2007, Chapter 10). It has the convenient property of being invariant under reparametrizations. This criterion is actually a particular case of D_S-optimality, in which only the parameters corresponding to a subset S of the indices are of interest. See Section 10.3 of Atkinson, Donev and Tobias (2007). Much of our theory and discussion will generalize to a broader class of criteria that we discuss in Section 3.2.

Before proceeding to the setting in which X and Z are random, we note a helpful result. Under the model (2), there is a convenient property of D-optimality in this setting, which we state as the following simple theorem.

Theorem 1. *For data following model* (2), *assume that U is full rank. Then the D-optimality criteria for $\tilde{\gamma}$, $\tilde{\beta}$ and δ are equivalent.*

Proof. D-optimality for $\tilde{\gamma}$ comes from minimizing $\det((U^\top U)^{-1})_{22})$ while D-optimality for δ comes from maximizing $\det(U^\top U)$. To show that those are equivalent, we write

$$U^\top U = \begin{bmatrix} \tilde{X}^\top \tilde{X} & \tilde{X}^\top D\tilde{X} \\ \tilde{X}^\top D\tilde{X} & \tilde{X}^\top D^2 \tilde{X} \end{bmatrix} = \begin{bmatrix} \tilde{X}^\top \tilde{X} & \tilde{X}^\top D\tilde{X} \\ \tilde{X}^\top D\tilde{X} & \tilde{X}^\top \tilde{X} \end{bmatrix} \equiv \begin{bmatrix} A & B \\ B & A \end{bmatrix} \qquad (4)$$

using $D^2 = I$ in the second equality. In the decomposition above, A and B are symmetric with A invertible and so, from properties of block matrices

$$\det(U^\top U) = \det(A)\det(A - BA^{-1}B) \quad \text{and}$$
$$((U^\top U)^{-1})_{22} = (A - BA^{-1}B)^{-1},$$

from which

$$\det\left(U^\top U\right) = \frac{\det(\tilde{X}^\top \tilde{X})}{\det(((U^\top U)^{-1})_{22})}. \qquad (5)$$

Because $\tilde{X}^\top \tilde{X}$ does not depend on Z, our D-optimality criterion is equivalent to maximizing $\det(U^\top U)$. Finally, D-optimality for $\tilde{\beta}$ and $\tilde{\gamma}$ are equivalent by symmetry. $\qquad\square$

The equivalence (5) is well-known in the D$_S$-optimality literature (see, e.g., Equation 10.7 of Atkinson, Donev and Tobias (2007)). In our setting, only the right half of the columns of U can be changed. Then the numerator in Equation (5) is conveniently fixed and so optimizing the denominator alone optimizes the ratio.

The simple structure of the model (2) has made *D*-optimal estimation of $\tilde{\gamma}$ equivalent to D-optimality for δ and for $\tilde{\beta}$. By the same token, a design that is A-optimal for $\tilde{\gamma}$, minimizing $\mathrm{tr}(\mathrm{Var}(\hat{\tilde{\gamma}}))$ is also A-optimal for δ and for $\tilde{\beta}$. Lemma 1 of Nachtsheim (1989) includes our setting and shows that *D*-optimality for δ is equivalent to *D*-optimality for $\tilde{\gamma}$. It does not apply to $\tilde{\beta}$ for our problem nor does that result consider other criteria such as A-optimality.

The theory of marginally restricted D-optimality in Cook and Thibodeau (1980) describes some settings where the D-optimal design for all variables is a tensor product of the given empirical design for the fixed variables and a randomized design for the settable variables. Such a design is simply an RCT on the settable variables. By their Lemma 1, this holds when the regression model is a Kronecker product of functions of fixed variables times functions of settable variables. In their Lemma 3, this holds when the regression model has an intercept plus a sum of functions of settable variables and a sum of functions of fixed variables. Neither of those apply to model (2) but Lemma 2 of Nachtsheim (1989) does. The TBD designs we consider usually have constraints on the short-term gain or monotonicity constraints, and those generally make RCTs non-optimal.

3.2. Prospective D-optimality

In this section, we modify the approach in Section 3.1 to account for the randomness in X and Z. To do so, we adopt a prospective D-optimality criterion to apply to the setting where both X_i and Z_i have yet to be observed. We derive our approach from ideas in Bayesian optimal design, which we briefly summarize below based on Chapter 18.2 of Atkinson, Donev and Tobias (2007), omitting details that will not play a role in our setting.

Bayesian optimal design often arises when the variance of the parameter estimate depends on the unknown true value of the parameter θ, as is the usual case for models where the expected response is a nonlinear function of θ. In this case, the information matrix $M = M(\theta)$ is usually constructed by linearizing the model form and then taking the expected outer product of the gradient with itself under a design measure on the predictors. This $M(\theta)$ is random because θ has a prior distribution. In our case, $M = U^\top U$ since this is the information matrix for the multivariate regression.

The favored approach in Bayesian optimal design is to choose the design in order to minimize $\mathbb{E}(\log \det(M^{-1}))$ where the expectation is over the prior distribution on the parameters (Chaloner and Verdinelli, 1995; Dette, 1996). That can be quite expensive to do. Many of the examples in the literature optimize the design over a grid which is reasonable when the dimensions of X and θ are both small, but those methods do not scale well to larger problems.

Table 18.1 of Atkinson, Donev and Tobias (2007) lists four additional criteria along with the above one. They are $\log \mathbb{E}(\det(M^{-1}))$, $\log \det(\mathbb{E}(M^{-1}))$, $\log(\mathbb{E}(\det(M))^{-1})$, and $\log(\det \mathbb{E}(M))^{-1}$. The objective becomes more tractable each time a nonlinear operation is taken out of the expectation. When the logarithm is the final step, the criterion is equivalent to not taking that logarithm.

We choose the last of those four quantities (choice V in their Table 18.1) for our definition of prospective D-optimality.

Definition 1. [Prospective D-optimality] For random predictors X, a design function $\mathbb{P}(Z = 1 \,|\, x) = p(x)$ is *prospectively D-optimal* if it maximizes

$$\det(\mathbb{E}(U^\top U)) = \det(\mathbb{E}[\text{Var}(\hat{\delta} \,|\, X, Z)^{-1}]),$$

where the expectation is with respect to X and Z.

We could analogously define prospective D-optimality for $\tilde{\beta}$ or $\tilde{\gamma}$ as minimizing $\det((\mathbb{E}[U^\top U]^{-1})_{11})$ or $\det((\mathbb{E}[U^\top U]^{-1})_{22})$, respectively. By Theorem 1, prospective D-optimality in the sense of Definition 1 is equivalent to these conditions in our model, so the three notions all align.

Our choice is known as EW D-optimality in Bayesian design for generalized linear models (GLMs) (Yang, Mandal and Majumdar, 2016; Bu, Majumdar and Yang, 2020). The E is for expectation and the W refers to a weight matrix arising in GLMs. It is valued for its significantly reduced computational cost. While a computationally efficient design is not necessarily the best one, Yang, Mandal and Majumdar (2016) and Yang, Tong and Mandal (2017) find in a range of

simulation studies that EW D-optimal designs tend to have strong performance under the more standard Bayesian D-optimality criterion as well.

Since our prospective D-optimality criterion uses the expected value of $U^\top U$, it depends only on the p_i terms and not on the joint distribution of the Z_i. It is therefore important that the Z_i be independent in order to distinguish, say, an RCT with $\mathbb{P}(Z_i = 1) = 1/2$ from an allocation that takes all $Z_i = Z_1$ where $Z_1 = 1$ with probability $1/2$. In addition, Morrison et al. (2024) show in a similar problem that, when the Z_i are sampled independently, the design that minimizes $\det(\mathbb{E}[M])$ also minimizes $\mathbb{E}[\det(M)]$ asymptotically as $n \to \infty$. Kluger and Owen (2023) consider some stratified sampling methods that incorporate negative correlations among the Z_i which makes $U^\top U$ come even closer to its expectation than it does under independent sampling.

Under sampling with $X_i \sim P_X$, define

$$
\begin{aligned}
\tilde{\Sigma} &= \mathbb{E}\Big(\frac{1}{n}\tilde{X}^\top \tilde{X}\Big) \\
&= \begin{bmatrix} 1 & \mathbb{E}[X_\bullet]^\top \\ \mathbb{E}[X_\bullet] & \mathbb{E}[X_\bullet X_\bullet^\top] \end{bmatrix},
\end{aligned}
$$

where the bullet subscript denotes an arbitrary subject with $X_\bullet \sim P_X$ and $\mathbb{P}(Z_\bullet = 1 \mid X_\bullet) = p(X_\bullet)$. In addition,

$$
\begin{aligned}
\mathbb{E}\Big(\frac{1}{n}(\tilde{X}^\top D \tilde{X})_{jk}\Big) &= \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} Z_i \tilde{X}_{ij}\tilde{X}_{ik}\Big) \\
&= \mathbb{E}[\tilde{X}_{\bullet j}\tilde{X}_{\bullet k}(2p(X_\bullet) - 1)]
\end{aligned}
$$

Now let N be the matrix with

$$
N_{jk} = \mathbb{E}[\tilde{X}_{\bullet j}\tilde{X}_{\bullet k}(2p(X_\bullet) - 1)]. \tag{6}
$$

Under our sampling assumptions

$$
\mathbb{E}\Big(\frac{1}{n}U^\top U\Big) = \begin{bmatrix} \tilde{\Sigma} & N \\ N & \tilde{\Sigma} \end{bmatrix}. \tag{7}
$$

The right-hand side of (7) represents the expected information per observation in our tie-breaker design. Using these formulas, we obtain the following desirable result for prospective D-optimality.

Theorem 2. *Under the model* (2) *with $X_i \sim P_X$, the RCT $p(X_\bullet) = 1/2$ is prospectively D-optimal among all designs $p(X_\bullet)$. Moreover, this is the unique D-optimal design of the form* (3).

The proofs of Theorem 2 and of all subsequent theorems are presented in Appendix A. Theorem 2 does not require X_i to be independent though that would be the usual model.

The theorem establishes that the RCT is prospectively D-optimal among *any* randomization scheme $\mathbb{P}(Z = 1 \mid X_\bullet) = p(X_\bullet) \in [0,1]$. It is not necessarily the

unique optimum in this larger class. For instance, if

$$\mathbb{E}[X_{\bullet j} X_{\bullet k}(2p(X_\bullet) - 1)] = 0$$

for all j and k, then the function $p(\cdot)$ would provide the same efficiency as an RCT since it would make the matrix N in the above proof vanish.

Though we frame Theorem 2 via prospective D-optimality, the result also holds for a broader class of criteria that we call *prospectively monotone*. The basic idea of such criteria is that they only depend on the bottom-right submatrix of $\mathbb{E}[(U^\top U)^{-1}]$ and that they encourage this submatrix to be small in the standard ordering on positive semi-definite matrices (i.e., $\Sigma_1 \preceq \Sigma_2$ if and only if $\Sigma_2 - \Sigma_1$ is PSD). The precise definition is as follows.

Definition 2. [Prospective monotonicity] For random predictors, we say that a criterion is *prospectively monotone* for $\hat{\delta}$ if it depends only on the matrix $\mathbb{E}[\mathrm{Var}(\hat{\hat{\gamma}} \,|\, X, Z)^{-1}]$ and if the criterion increases whenever this matrix increases in the standard ordering on positive semi-definite matrices.

Because these are the only properties of prospective D-optimality we use in Theorem 2, the same result holds immediately for any prospectively monotone criterion. Examples include prospective A-optimality, which minimizes the quantity $\mathrm{Tr}(\mathbb{E}[\mathrm{Var}(\hat{\hat{\gamma}} \,|\, Z)^{-1}])$, or prospective C-optimality, which minimizes $c^\top \mathbb{E}[\mathrm{Var}(\hat{\hat{\gamma}} \,|\, Z)^{-1}]c$ for some preset vector c (the latter of use if a particular linear combination of elements of $\hat{\gamma}$ is of primary interest).

3.3. Symmetric distributions with symmetric designs

We can gain particular insights, both theoretical and practical, by considering a special case satisfying two conditions. First, P_X has a symmetric density, i.e., $f_X(\vec{x}) = f_X(-\vec{x})$ for $\vec{x} \in \mathbb{R}^d$. This includes the special case of Gaussian covariates, which we consider in more detail as well. Secondly, we will further assume that $p = 1/2$ and the randomization window is symmetric about zero with width $\Delta \geqslant 0$, which we call a symmetric design. That is, we restrict (3) to simply

$$p(Z_i = 1 \mid X_i) = \begin{cases} 1, & X_i^\top \eta \geqslant \Delta \\ 1/2, & |X_i^\top \eta| \in (-\Delta, \Delta) \\ 0, & X_i^\top \eta \leqslant -\Delta. \end{cases} \tag{8}$$

For $j = k = 0$ (i.e., both terms are intercepts), equation (6) reduces to

$$N_{00} = \mathbb{E}[(\mathbb{1}\{X_\bullet^\top \eta \geqslant \Delta\} - \mathbb{1}\{X_\bullet^\top \eta \leqslant -\Delta\})] = 0$$

since we are integrating an odd function with respect to a symmetric density. Likewise, when both $j, k \geqslant 1$ we have $N_{jk} = 0$. The only cases that remain are the first row and first column of N, besides the top-left entry. Thus, we can write

$$N = \begin{bmatrix} 0 & \alpha^\top \\ \alpha & \mathbf{0}_{d \times d} \end{bmatrix} \tag{9}$$

where $\alpha \in \mathbb{R}^d$ with

$$\alpha_j = \mathbb{E}\big[X_{\bullet j}\big(\mathbb{1}\{X_{\bullet}^{\top}\eta \geqslant \Delta\} - \mathbb{1}\{X_{\bullet}^{\top}\eta \leqslant -\Delta\}\big)\big]$$
$$= 2\mathbb{E}[X_{\bullet j}\mathbb{1}\{X_{\bullet}^{\top}\eta \geqslant \Delta\}]. \tag{10}$$

We note that $\alpha = \alpha(\Delta, \eta)$ depends on the width Δ and the treatment assignment vector η, but we suppress that dependence for notational ease. From (10), we can compute explicitly that

$$N\tilde{\Sigma}^{-1}N = \begin{bmatrix} \alpha^{\top}\Sigma^{-1}\alpha & 0 \\ 0 & \alpha\alpha^{\top} \end{bmatrix},$$

so our criterion becomes

$$\det(\tilde{\Sigma})\det(\tilde{\Sigma} - N\tilde{\Sigma}^{-1}N) = \det(\Sigma)(1 - \alpha^{\top}\Sigma^{-1}\alpha)\det(\Sigma - \alpha\alpha^{\top})$$
$$= (1 - \alpha^{\top}\Sigma^{-1}\alpha)^2 \det(\Sigma)^2.$$

In the last line we use the formula $\det(A + cd^{\top}) = \det(A)(1 + d^{\top}A^{-1}c)$ for the determinant of a rank-one update of an invertible matrix and we also note that $\det(\tilde{\Sigma}) = \det(\Sigma)$. Let $W = \Sigma^{1/2}$ so that $\mathrm{Var}(W^{-1}x) = I$. The efficiency therefore only depends on α through $\alpha^{\top}\Sigma^{-1}\alpha = \|W^{-1}\alpha\|^2$.

We could also ask whether we can do better by changing our randomization scheme to allow

$$\mathbb{P}(Z_i = 1 \,|\, X_i) = \begin{cases} 1, & X_i^{\top}\eta \geqslant \Delta \\ p, & |X_i^{\top}\eta| < \Delta \\ 0, & X_i^{\top}\eta \leqslant -\Delta \end{cases} \tag{11}$$

for some other $p \neq 1/2$. While this may be a reasonable choice in practice when treatment cannot be assigned equally, it cannot provide any efficiency benefit in the symmetric case, as shown in Theorem 3 below. Just as an RCT is most efficient globally, if one is using the three level rule (11) then the best choice for the middle level is $1/2$ and that choice is unique under a reasonable assumption.

Theorem 3. *If P_X is symmetric and the randomization window is symmetric around zero with width Δ, then a prospectively D-optimal design of the form* (11) *has $p = 1/2$. Moreover, this design is unique provided that $\mathbb{P}(|X_{\bullet}^{\top}\eta| \leqslant \Delta) > 0$.*

An informative example is the case in which $P_X = \mathcal{N}(0, \Sigma)$ for some covariance matrix Σ. In this case, we can compute the efficiency explicitly as a function of Δ.

Theorem 4. *Let P_X be the $\mathcal{N}(0, \Sigma)$ distribution for a positive definite matrix Σ. For $X_i \overset{\mathrm{iid}}{\sim} P_X$ and Z_i sampled independently from* (3) *for a nonzero vector $\eta \in \mathbb{R}^d$ and a threshold $\Delta \geqslant 0$*

$$\det\big(\mathbb{E}\big(\mathrm{Var}(\hat{\delta}\,|\,Z)\big)^{-1}\big) = \Big(1 - \frac{2}{\pi}e^{\frac{-\Delta^2}{\eta^{\top}\Sigma\eta}}\Big)^2 \det(\Sigma)^2.$$

From Theorem [4] we find that the efficiency ratio between $\Delta = \infty$ (the RCT) and $\Delta = 0$ (the RDD) is $(1 - 2/\pi)^{-2} \approx 7.57$. The result in Goldberger (1972) gives a ratio of $(1 - 2/\pi)^{-1}$ for the variance of the slope in the case $d = 1$. Our result is the same, though we pick up an extra factor because our determinant criterion incorporates both the intercept and the slope. Their result was for $d = 1$; here we get the same efficiency ratio for all $d \geqslant 1$.

In this multivariate setting we see that for any fixed $\Delta > 0$, the most efficient design minimizes $\eta^\top \Sigma \eta$, so it is the eigenvector corresponding to the smallest eigenvalue of Σ. This represents the least "distribution-aware" choice, i.e., the last principal component vector, which aligns with our intuition that we gain more information by randomizing as much as possible.

4. Short-term gain

We turn now to the other arm of the tradeoff, the short-term gain. In our motivating problems, the response is defined so that larger values of Y_i define better outcomes. Now $\mathbb{E}[Y_i] = \mathbb{E}[\tilde{X}_i^\top \tilde{\beta}] + \tilde{T}_i$ where $\tilde{T}_i = \mathbb{E}[Z_i \tilde{X}_i^\top \tilde{\gamma}]$. The first term in $\mathbb{E}[Y_i]$ is not affected by the treatment allocation and so any consideration of short-term gain can be expressed in terms of $\tilde{T}_i$. Further, $\mathbb{E}(\tilde{T}_i) = \mathbb{E}[Z_i \gamma_0] + T_i$ for $T_i = \mathbb{E}[Z_i X_i^\top \gamma]$. Now $\mathbb{E}[Z_i \gamma_0]$ only depends on our design via the expected proportion of treated subjects. This will often be fixed by a budget constraint and even when it is not fixed, it does not depend on where specifically we assign the treatment. Therefore, for design purposes we may focus on T_i.

Under the model (2) with treatment assignment from (11),

$$T_i = \mathbb{E}\big[X_i^\top \gamma \big(\mathbb{1}\{X_i^\top \eta \geqslant u\} - \mathbb{1}\{X_i^\top \eta \leqslant \ell\} + (2p-1)\mathbb{1}\{X_i^\top \eta \in (\ell, u)\}\big)\big]. \quad (12)$$

If $\eta = \gamma$, so that we assign treatment using the true treatment effect vector, then equation (12) shows that the best expected gain comes from taking $u = \ell = 0$, which is an RDD centered at the mean. Moreover, the expected gain decreases monotonically as we increase u beyond 0 or decrease ℓ below 0. This matches our intuition that we must sacrifice some short-term gain to improve on statistical efficiency. Ordinarily $\eta \neq \gamma$, and a poor choice of η could break this monotonicity.

In the Gaussian case considered in Section 3.3, we can likewise derive an explicit formula for the expected gain as a function of u, η, and γ. Letting $T_\bullet = Z_\bullet X_\bullet^\top \gamma$, we have

$$
\begin{aligned}
\mathbb{E}[T_\bullet] &= \mathbb{E}[X_\bullet^\top \gamma (\mathbb{1}\{X_\bullet^\top \eta \geqslant u\} - \mathbb{1}\{X_\bullet^\top \eta \leqslant -u\})] \\
&= \gamma^\top \mathbb{E}[X_\bullet (\mathbb{1}\{X^\top \eta \geqslant u\} - \mathbb{1}\{X_\bullet^\top \eta \leqslant -u\})] \\
&= \gamma^\top \alpha.
\end{aligned}
$$

Using the formula (23) for α in the Gaussian case, this is simply

$$\mathbb{E}[T_\bullet] = \sqrt{\frac{2}{\pi}} \frac{\gamma^\top \Sigma \eta}{\sqrt{\eta^\top \Sigma \eta}} \, e^{\frac{-u^2}{2\eta^\top \Sigma \eta}}.$$

We expect intuitively that $\eta = \gamma$ will maximize $\mathbb{E}[T_\bullet]$. To verify this we start by choosing $u = u(\eta)$ in a way that keeps the proportion of data in the three treatment regions constant. We do so by taking $u = u(\eta) = u_0 \sqrt{\eta^\top \Sigma \eta}$ for some $u_0 \geqslant 0$, and then

$$\mathbb{E}[T_\bullet] = \sqrt{\frac{2}{\pi}} \frac{\gamma^\top \Sigma \eta}{\sqrt{\eta^\top \Sigma \eta}} \; e^{-u_0^2/2}.$$

Let $\gamma_w = \Sigma^{1/2} \gamma$ and $\eta_w = \Sigma^{1/2} \eta$, using the same matrix square root in both cases. Then

$$\frac{\gamma^\top \Sigma \eta}{\sqrt{\eta^\top \Sigma \eta}} = \frac{\gamma_w^\top \eta_w}{\|\eta_w\|}$$

is maximized by taking $\eta_w = \gamma_w$ or equivalently $\eta = \gamma$. Any scaling of $\eta = \gamma$ leaves this criterion invariant.

Working under the normalization $\eta^\top \Sigma \eta = 1$, we can summarize our results in the Gaussian case as

$$\det\left(\frac{1}{n} U^\top U\right) = \left(1 - \frac{2}{\pi} e^{-u_0^2}\right)^2 \det(\Sigma)^2, \quad \text{and} \tag{13}$$

$$\mathbb{E}[T_\bullet] = \sqrt{\frac{2}{\pi}} \gamma^\top \Sigma \eta \; e^{-u_0^2/2}. \tag{14}$$

With our normalization, $u^2 = u_0^2 \eta^\top \Sigma \eta = u_0^2$. Equations (13) and (14) quantify the tradeoff between efficiency and short-term gain, that come from choosing u_0. Greater randomization through larger u_0 increases efficiency, and, assuming that the sign of η is properly chosen, decreases the short-term gain.

5. Convex formulation

In this section, we return to the setting where $x_1, \ldots, x_n \in \mathbb{R}^d$ are fixed values but Z_i are not yet assigned. We use lowercase x_i to emphasize that they are non-random. We assume that any subset of the x_i of size d has full rank, as would be the case almost surely when they are drawn IID from a distribution whose covariance matrix is of full rank. The design problem is to choose $p_i = \mathbb{P}(Z_i = 1)$.

For given x_i, the design matrix in (2) is

$$U = \begin{bmatrix} u_1(Z_1)^\top \\ u_2(Z_2)^\top \\ \vdots \\ u_n(Z_n)^\top \end{bmatrix}, \quad \text{for} \quad u_i(1) = u_{i+} \equiv \begin{bmatrix} \tilde{x}_i \\ \tilde{x}_i \end{bmatrix} \quad \text{and} \quad u_i(-1) = u_{i-} \equiv \begin{bmatrix} \tilde{x}_i \\ -\tilde{x}_i \end{bmatrix}.$$

Introducing $p_{i+} = p_i$ and $p_{i-} = 1 - p_i$ we get

$$\mathbb{E}(U^\top U) = \sum_{i=1}^n \mathbb{E}(u_i(Z_i) u_i(Z_i)^\top) = \sum_{i=1}^n (p_{i+} u_{i+} u_{i+}^\top + p_{i-} u_{i-} u_{i-}^\top). \tag{15}$$

Our design criterion is to choose $p_{i\pm}$ to minimize

$$-\log\det(\mathbb{E}(U^\top U)) = -\log\det\left(\sum_{i=1}^n \sum_{s\in\{+,-\}} p_{is} u_{is} u_{is}^\top\right).$$

This problem is only well-defined for $n \geqslant d$, since otherwise the matrix does not have full rank for any choice of $p \in \mathbb{R}^n$ and the determinant is always zero. This criterion is convex in $\{p_{is} \mid 1 \leqslant i \leqslant n, s \in \{+,-\}\}$ by a direct match with Chapter 7.5.2 of Boyd and Vandenberghe (2004) over the convex domain with $0 \leqslant p_{i\pm} \leqslant 1$ and $p_{i+} + p_{i-} = 1$ for all i.

It will be simpler for us to optimize over $q \equiv 2p - 1 \in [-1, 1]^n$, in which case

$$-\log\det(\mathbb{E}(U^\top U)) = -\log\det\left(\begin{bmatrix} \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top & \sum_{i=1}^n q_i \tilde{x}_i \tilde{x}_i^\top \\ \sum_{i=1}^n q_i \tilde{x}_i \tilde{x}_i^\top & \sum_{i=1}^n \tilde{x}_i \tilde{x}_i^\top \end{bmatrix}\right). \tag{16}$$

Absent any other constraints, we have seen that the RCT ($q_i = 0$, $p_i = 1/2$ for all $i \leqslant n$) always minimizes (16). The constrained optimization of (16) can be cast as both a semi-definite program (Boyd and Vandenberghe, 2004) and a mixed integer second-order cone program (Sagnol and Harman, 2015).

This setting is close to the usual design measure relaxation. Instead of choosing $n_i = 1$ point $(\tilde{x}_i, Z_i)$ for observation i we make a random choice between $(\tilde{x}_i, 1)$ and $(\tilde{x}_i, -1)$ for that point. The difference here is that we have the union of n such tiny design problems.

5.1. Some useful convex constraints

We outline a few reasonable (and convex) constraints that one could impose on this problem:

Budget constraint: In practice we have a fixed budget for the treatments. For instance the number of scholarships or customer perks to give out may be fixed for economic reasons. We can impose this constraint in expectation by setting $(1/n)\sum_{i=1}^n p_i = \mu$, where μ is some fixed average treatment rate.

Monotonicity constraint: It may be reasonable to require that p_i is nondecreasing in some running variable $r_i = f(\tilde{x}_i)$. For example, a university may require that an applicant's probability of receiving a scholarship can only stay constant or increase with their score, which is some combination of applicant variables. We can encode this as a convex constraint by first permuting the data matrix so that $r_{(1)} \leqslant r_{(2)} \leqslant \cdots \leqslant r_{(n)}$ and then forcing $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(n)}$. Note that the formulation (3) satisfies this monotonicity constraint, in which case $r_i = \tilde{x}_i^\top \eta$.

Gain constraint: One may also want to impose that the expected gain is at least some fraction of its highest possible value, i.e.

$$\sum_{i=1}^n p_i \tilde{x}_i^\top \eta \geqslant \rho \sum_{i=1}^n |\tilde{x}_i^\top \eta|. \tag{17}$$

The left-hand side of (17) is the expected gain for this choice of p_i, whereas the right-hand side is the highest possible gain, which corresponds to the RDD $\mathbb{P}(Z_i = 1) = \mathbb{1}\{\tilde{x}_i^\top \eta \geqslant 0\}$. Because γ is typically not known exactly, (17) computes the anticipated gain under the sampling direction η we use. If an analyst has access to a better estimate of gain, such as a set of estimated treatment effects $\{\hat{\tau}(\tilde{x}_1), \ldots, \hat{\tau}(\tilde{x}_n)\}$ fit from prior data, they could replace (17) by $\sum_{i=1}^n p_i \hat{\tau}(\tilde{x}_i) \geqslant \rho \sum_{i=1}^n |\hat{\tau}(\tilde{x}_i)|$, which is again a linear constraint on q.

Covariate balance constraint: The expected total value of the jth variable for treated subjects is $\sum_i p_i \tilde{x}_{ij}$. The constraint

$$\sum_i p_i(\tilde{x}_{ij} - \Delta_j) \leqslant 0$$

keeps the expected sample average of the jth variable among treated subjects to at most Δ_j regardless of the actual number of treated subjects. We can similarly impose a lower bound on that expected sample average. We note that sufficiently stringent covariate balance constraints may not be compatible with gain or monotonicity constraints.

5.2. Piecewise constant optimal designs

Though we delay discussion of a particular example to Section 6, we prove here an empirically-observed phenomenon regarding the monotonicity constraint. In particular, optimal solutions when a monotonicity constraint is imposed display only a few distinct levels in the assigned treatment probability. This is consistent with the one-dimensional theory of Li and Owen (2023), though interestingly that paper observed the same behavior even with only a budget constraint. It also used an approach that does not obviously extend to $d > 1$.

Theorem 5. *Consider the optimization problem*

$$\begin{aligned}
\min_q \ & -\log\det\left(\sum_{i=1}^n \begin{bmatrix} \tilde{x}_i\tilde{x}_i^\top & q_i\tilde{x}_i\tilde{x}_i^\top \\ q_i\tilde{x}_i\tilde{x}_i^\top & \tilde{x}_i\tilde{x}_i^\top \end{bmatrix}\right) \\
\text{s.t.} \ & q \in [-1, 1]^n, \\
& q_{\pi(1)} \leqslant q_{\pi(2)} \leqslant \cdots \leqslant q_{\pi(n)}, \\
& \overline{q} = \mu.
\end{aligned} \tag{18}$$

Suppose that the non-intercept columns $x_1, \ldots, x_n$ are drawn IID from a density on $\mathbb{R}^d$. Then with probability one, the solution to (18) has at most $\binom{d+2}{2} + 2$ distinct levels.

These upper bounds closely resemble those in Cook and Fedorov (1995), who study constraints on a design of $\tilde{x}_1, \ldots, \tilde{x}_n$ themselves rather than on a treatment variable. However, theirs is an existence result for *some* solution with this level of sparsity, whereas our result holds for *any* solution with probability one. Moreover, their result derives from Caratheodory's theorem, whereas ours comes from a close analysis of the KKT conditions.

The upper bounds in Theorem 5 are not tight. In our experience, there have typically been no more than five or six distinct levels even for d as high as ten.

6. MIMIC-IV-ED example

In this section we detail a simulation based on a real data set of emergency department (ED) patients. The MIMIC-IV-ED database (Johnson et al., 2021) provided via PhysioNet (Goldberger et al., 2000) includes data on ED admissions at the Beth Israel Deaconess Medical Center between 2011 and 2019.

Emergency departments face heavy resource constraints, particularly in the limited human attention and beds available. It is thus important to ensure patients are triaged appropriately so that the patients in most urgent need of care are assigned to intensive care units (ICUs). In practice, this is often done via a scoring method such as the Emergency Severity Index (ESI), in which patients receive a score in $\{1, 2, 3, 4, 5\}$, with 1 indicating the highest severity and 5 indicating the lowest severity. MIMIC-IV-ED contains these values as acuity scores, along with a vector of vital signs and other relevant information about each patient.

Such a setting provides a very natural potential use case for tie-breaker designs. Patients arrive with an assortment of covariates, and hospitals acting under resource constraints must decide whether to put them in an ICU. A hospital or researcher may be interested in the treatment effect of an ICU bed; for example, a practical implication of such a question is whether to expand the ICU or allocate resources elsewhere (Phua, Hashmi and Haniffa, 2020). It is also of interest (Chang, Dacosta and Shapiro, 2017) to understand which types of patient benefit more or less from ICU beds to improve this costly resource allocation. Obviously, it is both unethical and counterproductive to assign some ICU beds by an RCT. Patients with high acuity scores must be sent to the ICU, and those with low acuity scores may be exposed unnecessarily to bad outcomes such as increased risk of acquiring a hospital-based infection (Kumar et al., 2018; Vranas et al., 2018). However, it may be possible to randomize "in the middle", e.g., by randomizing for patients with an intermediate acuity scores such as 3, or with scores in a range where there is uncertainty as to whether the ICU would be beneficial for that patient. Because such patients are believed to have similar severities, this would limit ethical concerns and allow for greater information gain.

The triage data set contains several vital signs for patients. Of these, we use all quantitative ones, which are: temperature, heart rate (HR), respiration rate (RR), oxygen saturation (O_2 Sat.), and systolic and diastolic blood pressure (SBP and DBP). There is also an acuity score for each patient, as described above. The data set contains 448,972 entries, but to arrive at a more realistic sample for a prospective analysis, we randomly select 200 subjects among those with no missing or blatantly inaccurate entries. Our optimization was done using the CVXR package (Fu, Balasubramanian and Boyd, 2020) and the MOSEK solver (ApS, 2022).
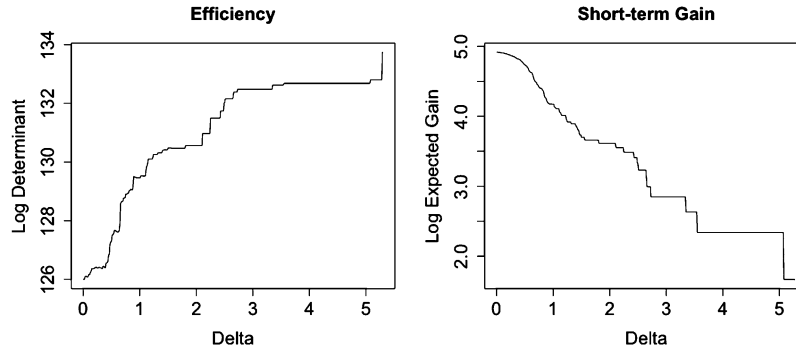
FIG 1. *Efficiency and gain of the MIMIC-IV-Ed simulation as a function of the width of the randomization window* Δ.

TABLE 1

These are the coefficients $\hat{\eta}_j$ that define a quadratic running variable for the MIMIC data. The intercept is followed by a sum of pure quadratics in temperature, heart rate, respiration rate, O_2 saturation, systolic blood pressure and diastolic blood pressure.

Int.	Temp.	Temp^2	HR	HR^2	RR	RR^2
-0.74	-0.32	0.22	-0.03	0.67	-0.03	0.54
	O_2 Sat.	O_2 Sat.^{2}	SBP	SBP^2	DBP	DBP^2
	0.03	0.36	0.01	0.17	-0.11	-0.13

To carry out a full analysis of the sort described in this paper, we need a vector η, as in (3). In practice, one could assume a model of the form (2) and take $\eta = \hat{\gamma}$ for some estimate of γ formed via prior data. Since we do not have any Y values (which in practice could be some measure of survival, length of stay, or subsequent readmission), we will construct η via the acuity scores, using the reasonable assumption that treatment benefit increases with more severe acuity scores.

We collapse acuity scores of $\{1, 2\}$ into a group ($Y = 1$) and acuity scores of $\{3, 4, 5\}$ into another ($Y = 0$) and perform a logistic regression using these binary groups. The covariates used are the vital signs and their squares, the latter to allow for non-monotonic effects, e.g., the acuity score might be lower for both abnormally low and abnormally high heart rates. All covariates were scaled to mean zero and variance one. For pure quadratic terms the squares of the scaled covariates were themselves scaled to have mean zero and variance one. We also considered an ordered categorical regression model but preferred the logistic regression for ease of interpretability. Our estimated $\hat{\eta}_j$ are in Table 1.

Figure 1 presents the efficiency/gain tradeoff as we vary the size of the randomization window Δ in (8). For ease of visualization, we plot the logs of both quantities. As expected, we get a clear monotone increase in efficiency and decrease in gain as we increase Δ, moving from an RDD to an RCT. It should be noted that our efficiency criterion, because it only uses information in the X values, is robust to a poor choice of η, whereas our gain definition is con-
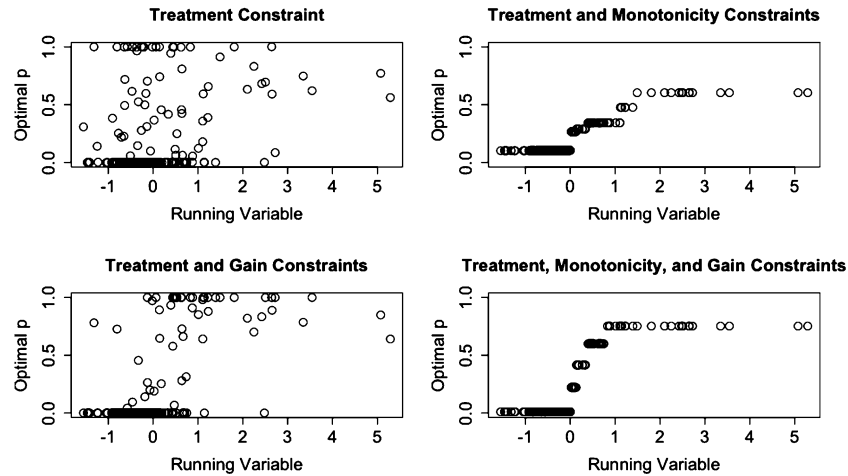
FIG 2. *Optimal solutions for MIMIC-IV-ED treatment probabilities under various constraints. The treatment constraint imposed $\overline{p} = 0.2$ for the average treatment rate, and the gain constraint imposed $\rho = 0.7$, i.e., at least 70% of the maximum possible gain.*

strained by the assumption that η is a reasonably accurate stand-in for the true treatment effect γ.

In practice, it is hard to interpret what a "good" value of efficiency is because of our D-optimality criterion. Hence, as in Owen and Varian (2020), a pragmatic approach is to first stipulate that the gain is at least some fraction of its highest possible value, and then pick the largest Δ for this choice to maximize efficiency. A more qualitative choice based on results like Figure 1, such as picking the right endpoint of a sharp efficiency jump or the left endpoint of a sharp gain decline, would also be sensible.

As we see in Figure 2, the treatment constraint causes most p_i to be at or near zero or one. Adding the gain constraint pushes most of the treatment probabilities to zero for low values of the running variable and one for high values. This scenario most closely resembles the RDD, with some deviations to boost efficiency. Indeed, the optimal solution would necessarily tend towards the RDD solution as the gain constraint increased. Finally, the monotonicity constraint further pushes the higher values of p to the positive values of the running variable and vice-versa, since we lose the opportunity to counterbalance some high and low probabilities at the extreme with their opposites. The right two panels display a few discrete levels in the treatment probability, consistent with Theorem 5.

7. Discussion

In this paper, we add to a growing body of work demonstrating the benefits of tie-breaker designs. Though RCTs are often infeasible, opportunities for small

doses of randomization may present themselves in a wide variety of real-world settings, in which case treatment effects can be learned more efficiently. This phenomenon is analogous to similar causal inference findings about merging observational and experimental data (Rosenman et al., 2021, 2023; Colnet et al., 2024).

The convex optimization framework in Section 5 is more general and conveniently only relies on knowing sample data rather than population parameters. It is also simple to implement and allows one to incorporate natural economic and ethical constraints with ease.

Multivariate tie-breaker designs are a natural option in situations in which there is no clear univariate running variable. For example, subjects may possess a vector of covariates, many of which could be associated with heterogeneous treatment effects in some unknown way of interest. Of course, two-line models and their multivariate analogs are not nearly as complicated as many of the models found in practice. Our view is to use them as a working model by which to decide on treatment allocations, in which case a more flexible model could be used upon full data acquisition as appropriate.

Appendix A: Proofs

Proof of Theorem 2. Using $\mathbb{E}(\mathrm{Var}(\hat{\delta}\,|\,X, Z)^{-1})$ from (7), we have

$$
\det \begin{bmatrix} \tilde{\Sigma} & N \\ N & \tilde{\Sigma} \end{bmatrix} = \det(\tilde{\Sigma}) \det(\tilde{\Sigma} - N\tilde{\Sigma}^{-1}N)
$$
$$
= \det(\tilde{\Sigma})^2 \det(I - \tilde{\Sigma}^{-1/2}N\tilde{\Sigma}^{-1}N\tilde{\Sigma}^{-1/2})
$$
$$
= \det(\tilde{\Sigma}^2) \det(I - A)
$$

where $A = \tilde{\Sigma}^{-1/2}N\tilde{\Sigma}^{-1}N\tilde{\Sigma}^{-1/2}$ is symmetric and positive semi-definite. Now $\det(I - A) \leqslant 1$ with equality if and only if $A = 0$, which occurs if and only if $N = 0$. Therefore, any design giving $N = 0$ is prospectively D-optimal. From the formula (6) for N_{jk}, we see right away that the RCT $p(X_\bullet) = 1/2$ satisfies this, which proves the first half of the theorem.

If we restrict to designs of the form (3), then for general (ℓ, u, p, η) we have

$$
N_{jk} = \mathbb{E}[\tilde{X}_{\bullet j}\tilde{X}_{\bullet k}(\mathbb{1}\{X_\bullet^\top \eta \geqslant u\} - \mathbb{1}\{X_\bullet^\top \eta \leqslant \ell\} + q\mathbb{1}\{X^\top \eta \in (\ell, u)\})]
$$

where $q \equiv 2p - 1$. Consider a tuple (ℓ, u, p, η) for which $N = 0$. Then considering in particular the entry of N where $j = k = 1$ (so that $\tilde{X}_{\bullet j} = \tilde{X}_{\bullet k} = 1$), we must have

$$
p_u - p_\ell + qp_m = 0 \tag{19}
$$

where $p_u = \mathbb{P}(X^\top \eta \geqslant u)$, $p_\ell = \mathbb{P}(X^\top \eta \leqslant \ell)$, and $p_m = \mathbb{P}(X^\top \eta \in (\ell, m))$. If we freeze $k = 0$ and vary j from 1 to d, we obtain

$$
\mathbb{E}[X_{\bullet j}(\mathbb{1}\{X_\bullet^\top \eta \geqslant u\} - \mathbb{1}\{X_\bullet^\top \eta \leqslant \ell\} + q\mathbb{1}\{X^\top \eta \in (\ell, u)\})] = 0
$$

for all j. Taking a suitable linear combination gives

$$\mathbb{E}[X_\bullet^\top \eta(\mathbb{1}\{X_\bullet^\top \eta \geqslant u\} - \mathbb{1}\{X_\bullet^\top \eta \leqslant \ell\} + q\mathbb{1}\{X^\top \eta \in (\ell, u)\})] = 0$$

Let $f(u) = \mathbb{E}[X^\top \eta \,|\, X^\top \eta \geqslant u]$, and analogously for $f(\ell)$ and $f(m)$. Then

$$f(u)p_u - f(\ell)p_\ell + f(m)qp_m = 0. \tag{20}$$

Next, equations (19) and (20) give respectively that

$$\begin{aligned} f(\ell)p_\ell &= f(\ell)p_u + f(\ell)qp_m \\ &= f(u)p_u + f(m)qp_m \end{aligned}$$

i.e., that

$$(f(u) - f(\ell))p_u = (f(\ell) - f(m))qp_m.$$

By definition, $f(u) > f(m) > f(\ell)$, so the left-hand side is non-negative and $(f(\ell) - f(m))$ is negative. This implies that $qp_m \leqslant 0$. Similarly, we have

$$f(u)p_u = f(u)p_\ell - f(u)qp_m = f(\ell)p_\ell - f(m)qp_m$$

and so

$$(f(u) - f(\ell))p_\ell = (f(u) - f(m))qp_m.$$

Therefore $qp_m \geqslant 0$, and so $qp_m = 0$. This leaves three possibilities: $q = 0$, $p_m = 0$, or both. If $p_m = 0$, then (19) implies that $p_u = p_\ell = 1/2$, which is impossible given (20). So then it must be that $q = 0$, whereupon (19) again implies that $p_u = p_\ell$ and (20) forces $p_u = p_\ell = 0$. In summary, a prospectively D-optimal design must satisfy

$$\mathbb{P}(X^\top \eta < \ell) = \mathbb{P}(X^\top \eta > u) = 0, \quad \text{and}$$
$$\mathbb{P}(Z = 1 \,|\, X^\top \eta \in (\ell, u)) = \frac{1}{2}. \qquad \square$$

Proof of Theorem 3. Let $q = 2p-1$. The off-diagonal block matrix $N = N(q)$ in (7) can now be written as

$$N_{jk} = \mathbb{E}[\tilde{X}_{\bullet j}\tilde{X}_{\bullet k}(\mathbb{1}\{X_\bullet^\top \eta \geqslant u\} - \mathbb{1}\{X_\bullet^\top \eta \leqslant u\} + q\mathbb{1}\{|X_\bullet^\top \eta| < u\})].$$

That is, we can write $N = N_0 + qN_1$, where N_0 is as in (9) and N_1 has (j, k) entry equal to $\mathbb{E}[\tilde{X}_{\bullet j}\tilde{X}_{\bullet k}\mathbb{1}\{|X_\bullet^\top \eta| < u\}]$. Note that N_1 is block diagonal, the exact opposite of N_0. Let

$$f(q) = \log\det\big(\tilde{\Sigma} - (N_0 + qN_1)\tilde{\Sigma}^{-1}(N_0 + qN_1)\big). \tag{21}$$

To prove the theorem, we will simply show that $f'(0) = 0$ and $f''(q) \leqslant 0$ for $q \in [-1, 1]$, implying that $q = 0$ (i.e., $p = 1/2$) is the global maximizer of f on this interval. Let

$$\begin{aligned} A &= -N_1\tilde{\Sigma}^{-1}N_1, \\ B &= -(N_1\tilde{\Sigma}^{-1}N_0 + N_0\tilde{\Sigma}^{-1}N_1), \quad \text{and} \\ C &= \tilde{\Sigma}^{-1} - N_0\tilde{\Sigma}^{-1}N_0 \end{aligned} \tag{22}$$

so that $f(q) = \log \det(q^2 A + qB + C)$. Call a $(d+1) \times (d+1)$ block matrix "block off-diagonal" if it is zero in the top-left entry and zero in the bottom-right $d \times d$ block, as in the case of N_0. The product of two block off-diagonal matrices is block-diagonal (with blocks of size 1×1 and $d \times d$), and the product of a block off-diagonal matrix and such a block diagonal matrix is block off-diagonal. Thus, A and C are both block diagonal, whereas B is block off-diagonal. Differentiating f, we obtain

$$f'(q) = \text{tr}((q^2 A + qB + C)^{-1}(2qA + B))$$

so that $f'(0) = \text{tr}(C^{-1}B)$. As noted, C is block diagonal and B is block off-diagonal, so the product $C^{-1}B$ is block off-diagonal and thus $f'(0) = 0$. It simplifies some expressions to let $M_1 = 2qA + B$ and $M_2 = (q^2 A + qB + C)^{-1}$. Then $f'(q) = \text{tr}(M_2 M_1)$ and

$$f''(q) = \text{tr}(-M_1 M_2 M_1 M_2 + 2M_2 A).$$

For $q \in [-1, 1]$, M_2 is the upper-left block of the inverse of the covariance matrix in (7), so it is positive semi-definite. Then $M_2^{1/2} M_1 M_2 M_1 M_2^{1/2}$ is positive semi-definite as well and thus

$$\text{tr}(-M_1 M_2 M_1 M_2) = -\text{tr}(M_2^{1/2} M_1 M_2 M_1 M_2^{1/2}) \leqslant 0.$$

In addition, A is negative semi-definite, so

$$\text{tr}(2M_2 A) = 2\text{tr}(M_2^{1/2} A M_2^{1/2}) \leqslant 0.$$

Therefore, $f''(q) \leqslant 0$ everywhere, so $q = 0$ is in fact a global optimum.

If $\mathbb{P}(|X_\bullet^\top \eta| \leqslant u) > 0$, then A and B are both nonzero, so these two trace inequalities are strict. Then $f''(q) < 0$ for all $q \in [-1, 1]$, so f cannot be constant anywhere. Since f is also concave on $[-1, 1]$, the local optimum at $q = 0$ must be a global optimum on this interval. $\square$

Proof of Theorem 4. To prove Theorem 4, we begin with two lemmas. We write φ for the $\mathcal{N}(0, 1)$ probability density function. We start our study of efficiency by finding an expression for α_j.

Lemma 1. *Let P_X be the $\mathcal{N}(0, I_d)$ distribution, let α_j be given by (10) and let Z_i be sampled according to (3) for a nonzero vector $\eta \in \mathbb{R}^d$ and $u \geqslant 0$. Then*

$$\alpha_j = 2\frac{\eta_j}{\|\eta\|}\varphi\Big(\frac{u}{\|\eta\|}\Big)$$

for $j = 1, \ldots, d$.

Proof. The result is easy if $\eta_j = 0$. Without loss of generality, assume that $\eta_j > 0$. Let x_{-j} and η_{-j} be the vectors in $\mathbb{R}^{d-1}$ formed by removing the jth component from the vectors x and η, respectively. Using $\varphi'(t) = -t\varphi(t)$,

$$\mathbb{E}[x_j(\mathbb{1}\{x^\top \eta \geqslant u\})] = \mathbb{E}\left[x_j \mathbb{1}\{x_j \geqslant (u - x_{-j}^\top \eta_{-j})/\eta_j\}\right]$$

$$= \mathbb{E}\left[\varphi((u - x_{-j}^{\top}\eta_{-j})/\eta_j)\right]$$

and applying it a second time along with symmetry of φ, we get

$$\alpha_j = \mathbb{E}[x_j(\mathbb{1}\{x^{\top}\eta \geqslant u\} - \mathbb{1}\{x^{\top}\eta \leqslant -u\})]$$
$$= \mathbb{E}\left[\varphi((u - x_{-j}^{\top}\eta_{-j})/\eta_j) + \varphi((u + x_{-j}^{\top}\eta_{-j})/\eta_j)\right].$$

Now let $\tilde{u}_j = u/\eta_j$ and $\tilde{z}_j = x_{-j}^{\top}\eta_{-j}/\eta_j \sim \mathcal{N}(0, \tau^2)$ with $\tau^2 = \|\eta_{-j}\|^2/\eta_j^2$. Then we get

$$\alpha_j = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2\pi\tau^2}} \int_{-\infty}^{\infty} \left(e^{-\frac{1}{2}(\tilde{u}_j - \tilde{z}_j)^2} + e^{-\frac{1}{2}(\tilde{u}_j + \tilde{z}_j)^2} \right) e^{-\tilde{z}_j^2/2\tau^2} \, d\tilde{z}_j$$
$$= \frac{1}{2\pi\sqrt{\tau^2}} \left(\frac{\sqrt{2\pi\tau^2}}{\sqrt{\tau^2 + 1}} e^{\frac{-\tilde{u}^2}{2(\tau^2 + 1)}} + \frac{\sqrt{2\pi\tau^2}}{\sqrt{\tau^2 + 1}} e^{\frac{-\tilde{u}^2}{2(\tau^2 + 1)}} \right)$$
$$= \sqrt{\frac{2}{\pi}} \frac{\eta_j}{\|\eta\|_2} e^{\frac{-u^2}{2\|\eta\|_2^2}}. \qquad \square$$

Lemma 2. *Let P_X be the $\mathcal{N}(0, \Sigma)$ distribution for a positive definite matrix Σ, let α_j be given by* (10) *and let Z_i be sampled according to* (3) *for a nonzero vector $\eta \in \mathbb{R}^d$ and $u \geqslant 0$. Then*

$$\alpha = \sqrt{\frac{2}{\pi}} \frac{\Sigma\eta}{\sqrt{\eta^{\top}\Sigma\eta}} e^{\frac{-u^2}{2\eta^{\top}\Sigma\eta}} = 2\frac{\Sigma\eta}{\sqrt{\eta^{\top}\Sigma\eta}} \varphi\left(\frac{u}{\sqrt{\eta^{\top}\Sigma\eta}}\right). \qquad (23)$$

Proof. For the general case of $x \sim \mathcal{N}(0, \Sigma)$, with Σ any positive-definite matrix, we define $W = \Sigma^{1/2}$ and write $x = Wz$. Then $z \sim \mathcal{N}(0, I_d)$, and so

$$\alpha_j = \mathbb{E}[x_j(\mathbb{1}\{x^{\top}\eta \geqslant u\} - \mathbb{1}\{x^{\top}\eta \leqslant -u\})]$$
$$= \mathbb{E}[W_j^{\top}(z\mathbb{1}\{z^{\top}W\eta \geqslant u\} - \mathbb{1}\{z^{\top}W\eta \leqslant -u\})]$$
$$= W_j^{\top}\mathbb{E}[z(\mathbb{1}\{z^{\top}W\eta \geqslant u\} - \mathbb{1}\{z^{\top}W\eta \leqslant -u\})].$$

This reduces the problem to the case $\Sigma = I_d$ with η replaced by $W\eta$, so we obtain

$$\alpha = W\left(\sqrt{\frac{2}{\pi}} \frac{W\eta}{\|W\eta\|_2} e^{\frac{-u^2}{2\|W\eta\|_2^2}}\right). \qquad \square$$

We can now conclude the proof of Theorem 4. By Lemma 2, we have

$$\alpha = \sqrt{\frac{2}{\pi}} \frac{\Sigma\eta}{\sqrt{\eta^{\top}\Sigma\eta}} e^{\frac{-u^2}{2\eta^{\top}\Sigma\eta}} = 2\frac{\Sigma\eta}{\sqrt{\eta^{\top}\Sigma\eta}} \varphi\left(\frac{u}{\sqrt{\eta^{\top}\Sigma\eta}}\right). \qquad (24)$$

Therefore,

$$\det\left(\mathbb{E}\left(\frac{1}{n}U^{\top}U\right)\right) = (1 - \alpha^{\top}\Sigma^{-1}\alpha)^2 \det(\Sigma)^2 = \left(1 - \frac{2}{\pi}e^{\frac{-u^2}{\eta^{\top}\Sigma\eta}}\right)^2 \det(\Sigma)^2. \qquad \square$$

Proof of Theorem 5. In the optimization problem (18), π is a fixed permutation in S_n corresponding to a monotonicity constraint, and so it is no loss of generality to take π to be the identity and omit it. If $\mu = -1$ or 1, then clearly the only solution is the constant vector $q = \mu$ (all treated or all control), so we ignore that case henceforth.

Because we aim to show that optimal solutions have constant levels in q, it will be easier to reparametrize in terms of the increments $c_i = q_i - q_{i-1}$ and show sparsity, where we take $c_1 = q_1$. Letting

$$G_0 = \begin{bmatrix} \tilde{X}^\top \tilde{X} & 0 \\ 0 & \tilde{X}^\top \tilde{X} \end{bmatrix}, \quad G_k = \sum_{i=k}^{n} \begin{bmatrix} 0 & \tilde{x}_i \tilde{x}_i^\top \\ \tilde{x}_i \tilde{x}_i^\top & 0 \end{bmatrix},$$

and $G(c) = G_0 + c_1 G_1 + \cdots + c_n G_n$, we can rewrite this problem as

$$\min_c \ -\log\det(G(c))$$

$$\text{such that } c_1 \geqslant -1,$$

$$c_k \geqslant 0 \text{ for all } 2 \leqslant k \leqslant n, \tag{25}$$

$$c^\top 1 \leqslant 1,$$

$$c_1 + \frac{n-1}{n} c_2 + \frac{n-2}{n} c_3 + \cdots + \frac{1}{n} c_n = \mu.$$

The first three constraints correspond to $q \in [-1, 1]^n$ and the monotonicity condition, while the last constraint corresponds to $\overline{q} = \mu$. Let v be the vector with $v_k = (n-k+1)/n$, so that the last constraint in (25) is $c^\top v = \mu$. Moreover, let

$$A = \begin{bmatrix} -1 & 0 & \ldots & 0 \\ 0 & -1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & -1 \\ 1 & 1 & \ldots & 1 \end{bmatrix} \in \mathbb{R}^{(n+1) \times n} \quad \text{and} \quad d = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^{n+1}$$

so that the inequality constraints can be written simply as $Ac - d \leqslant 0$. Slater's condition holds by considering the interior point

$$c = \left(\mu - (n-1)\epsilon, \frac{n}{n-1}\epsilon, \frac{n}{n-2}\epsilon, \ldots, \frac{n}{2}\epsilon, n\epsilon \right)$$

for any $\epsilon > 0$ sufficiently small, so the KKT conditions are both necessary and sufficient. The Lagrangian is

$$L(c; \lambda, \nu) = -\log\det(G(c)) + \sum_{i=1}^{n+1} \lambda_i (a_i^\top c - d_i) + \nu(c^\top v - \mu),$$

where a_i is a column vector formed from the i'th row of A. The differential is then

$$\frac{\partial L}{\partial c_k} = -\text{tr}(G(c)^{-1} G_k) + \sum_{i=1}^{n+1} \lambda_i a_{ik} + \nu v_k$$

$$= -2 \sum_{i=k}^{n} \tilde{x}_i^\top G(c)_{12}^{-1} \tilde{x}_i - \lambda_k + \lambda_{n+1} + \frac{n-k+1}{n} \nu.$$

Here, $G(c^*)_{12}^{-1} \in \mathbb{R}^{(d+1) \times (d+1)}$ is the off-diagonal block of $G(c^*)^{-1}$, which is symmetric since the off-diagonal block of $G(c^*)$ itself is symmetric.

The stationarity condition is that the partial derivative above is zero for all k. Suppose first that, for the optimal c^*, $\{c_1^*, c_2^*, \ldots, c_m^*\}$ are all nonzero, where m is thus far unspecified. Later, we relax the condition that it is the first m components of c specifically that are nonzero.

The complementary slackness condition implies that $c_k^* \lambda_k = 0$ for all $1 \leqslant k \leqslant n$, so in order for this to hold we must have $\lambda_1 = \lambda_2 = \cdots = \lambda_m = 0$. The stationarity condition for these terms is then

$$\sum_{i=k}^{n} \tilde{x}_i^\top G(c^*)_{12}^{-1} \tilde{x}_i = \lambda_{n+1}^* + \frac{n-k+1}{n} \nu^*, \quad \text{for all } 1 \leqslant k \leqslant m. \qquad (26)$$

The right-hand side of (26) is linear in k, with slope and intercept that depend on λ_{n+1}^* and ν^*. These constitute two free parameters on the right-hand side. For a fixed matrix $G(c)$, this is therefore a probability zero event whenever $m \geqslant 3$, since $x_1, \ldots, x_n$ are sampled IID from a density. However, the issue here is that c^* is not fixed but rather an implicit function of the data matrix X. Thus, we instead show that, for m sufficiently large and with probability one, there is *no* symmetric matrix $M \in \mathbb{R}^{(d+1) \times (d+1)}$ and pair (λ_{n+1}^*, ν^*) with

$$\sum_{i=k}^{n} \tilde{x}_i^\top M \tilde{x}_i = \lambda_{n+1}^* + \frac{n-k+1}{n} \nu^*, \quad \text{for all } 1 \leqslant k \leqslant m. \qquad (27)$$

Here, the logic is analogous: there are $\binom{d+2}{2}$ free parameters for an arbitrary symmetric matrix $M \in \mathbb{R}^{(d+1) \times (d+1)}$, as well as two additional parameters from λ_{n+1}^* and ν^*. Thus, if $m \geqslant \binom{d+2}{2} + 3$, this cannot happen with probability one. The logic is the same for any size m subset of $\{\tilde{x}_1, \ldots, \tilde{x}_n\}$, not just the first m, so the result follows by a union bound. $\qquad \square$

Acknowledgments

Funding

References

AIKEN, L., WEST, S., SCHWALM, D., CARROLL, J. and HSIUNG, C. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review* **22** 207–244.

ANGRIST, J., HUDSON, S. and PALLAIS, A. (2014). Leveling up: Early results from a randomized evaluation of post-secondary aid Technical Report, National Bureau of Economic Research.

APS, M. (2022). MOSEK Fusion API for Python 9.3.18. MR2323647

ATKINSON, A., DONEV, A. and TOBIAS, R. (2007). *Optimum Experimental Designs, with SAS. Oxford Statistical Science Series* **34**. Oxford University Press, Oxford.

BORUCH, R. F. (1975). Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods & Research* **4** 31–53. MR2061575

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization.* Cambridge University Press, Cambridge. MR4102684

BU, X., MAJUMDAR, D. and YANG, J. (2020). D-optimal designs for multinomial logistic models. *The Annals of Statistics* **48** 983–1000.

CAMPBELL, D. T. (1969). Reforms as experiments. *American Psychologist* **24** 409.

CAPPELLERI, J. C., DARLINGTON, R. B. and TROCHIM, W. M. K. (1994). Power analysis of cutoff-based randomized clinical trials. *Evaluation Review* **18** 141–152.

CATTANEO, M. D. and TITIUNIK, R. (2022). Regression discontinuity designs. *Annual Review of Economics* **14** 821–851. MR1390519

CHALONER, K. and VERDINELLI, I. (1995). Bayesian experimental design: A review. *Statistical Science* 273–304.

CHANG, D., DACOSTA, D. and SHAPIRO, M. (2017). Priority levels in medical intensive care at an academic public hospital. *JAMA Intern Med.* **177** 280–281. MR4718532

COLNET, B., MAYER, I., CHEN, G., DIENG, A., LI, R., VAROQUAUX, G., VERT, J. P., JOSSE, J. and YANG, S. (2024). Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science* **39** 165–191. MR1318209

COOK, D. and FEDOROV, V. (1995). Constrained optimization of experimental design. *Statistics* **26** 129–148. MR0577367

COOK, R. D. and THIBODEAU, L. (1980). Marginally restricted D-optimal designs. *Journal of the American Statistical Association* 366–371. MR1401846

DETTE, H. (1996). A note on Bayesian c- and D-optimal designs in nonlinear regression models. *The Annals of Statistics* **24** 1225–1234.

FU, A., BALASUBRAMANIAN, N. and BOYD, S. (2020). CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software* **94** 1–34. https://doi.org/10.18637/jss.v094.i14.

GELMAN, A., HILL, J. and VEHTARI, A. (2020). *Regression and Other Stories*. Cambridge University Press. MR3968384

GELMAN, A. and IMBENS, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics* **37** 447–456.

GOLDBERGER, A. (1972). Selection bias in evaluating treatment effects: Some formal illustrations. Technical Report Discussion paper, Institute for Research on Poverty, University of Wisconsin–Madison.

GOLDBERGER, A. L., AMARAL, L. A. N., GLASS, L., HAUSDORFF, J. M., IVANOV, P. C., MARK, R. G., MIETUS, J. E., MOODY, G. B., PENG, C. K. and STANLEY, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101** e215–e220.

HAHN, J., TODD, P. and DER KLAAUW, W. V. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

HEAVLIN, W. D. and FINNEGAN, G. P. (1998). Columnwise construction of response surface designs. *Statistica Sinica* 185–206.

JACOB, R., ZHU, P., SOMERS, M. A. and BLOOM, H. (2012). A practical guide to regression discontinuity. Working Paper, MDRC.

JOHNSON, A., BULGARELLI, L., POLLARD, T., CELI, L. A., MARK, R. and HORNG, S. (2021). MIMIC-IV-ED. https://doi.org/10.13026/77z6-9w59. MR4536118

KLUGER, D. M. and OWEN, A. B. (2023). Kernel regression analysis of tie-breaker designs. *Electronic Journal of Statistics* **17** 243–290.

KRANTZ, C. (2022). Modeling the outcomes of a longitudinal tie-breaker regression discontinuity design to assess an in-home training program for families at risk of child abuse and neglect, PhD thesis, University of Oklahoma.

KUMAR, S., SHANKAR, B., ARYA, S., DEB, M. and CHELLANI, H. (2018). Healthcare associated infections in neonatal intensive care unit and its correlation with environmental surveillance. *Journal of Infection and Public Health* **11** 275–279.

LI, H. and OWEN, A. B. (2022). A general characterization of optimality in tie-breaker designs. Technical Report No. arXiv2202.12511, Stanford University. MR4630939

LI, H. and OWEN, A. B. (2023). A general characterization of optimality in tie-breaker designs. *The Annals of Statistics* **51** 1030–1057.

LIPSEY, M. W., CORDRAY, D. S. and BERGER, D. E. (1981). Evaluation of a juvenile diversion program: Using multiple lines of evidence. *Evaluation Review* **5** 283–306. MR2109506

LOPEZ-FIDALGO, J. and GARCET-RODRIGUEZ, S. A. (2004). Optimal experimental designs when some independent variables are not subject to control. *Journal of the American Statistical Association* **99** 1190–1199. MR3718161

METELKINA, A. and PRONZATO, L. (2017). Information-regret compromise in covariate-adaptive treatment allocation. *The Annals of Statistics* **45**(5) 2046–2073.

Morrison, T., Nguyen, M., Baiocchi, M. and Owen, A. B. (2024). Constrained design of a binary instrument in a partially linear model. Technical Report, arXiv:2406.05592. MR1004348

Nachtsheim, C. J. (1989). On the design of experiments in the presence of fixed covariates. *Journal of Statistical Planning and Inference* **22** 203–212. MR4167610

Owen, A. B. and Varian, H. (2020). Optimizing the tie-breaker regression discontinuity design. *Electronic Journal of Statistics* **14** 4004–4027. `https://doi.org/10.1214/20-EJS1765`.

Phua, J., Hashmi, M. and Haniffa, R. (2020). ICU beds: less is more? Not sure. *Intensive Care Med.* **46** 1600–1602. MR4376789

Rosenman, E. T. R., Baiocchi, M., Banack, H. and Owen, A. B. (2021). Propensity score methods for merging observational and experimental datasets. *Statistics in Medicine.* MR4680697

Rosenman, E. T. R., Basse, G., Owen, A. B. and Baiocchi, M. (2023). Combining observational and experimental datasets using shrinkage estimators. *Biometrics* **79** 2961–2973. MR3396983

Sagnol, G. and Harman, R. (2015). Computing exact D-optimal designs by mixed integer second-order cone programming. *The Annals of Statistics* **43** 2198–2224.

Slivkins, A. (2019). Introduction to multi-armed bandits. *Foundations and Trends in Machine Learning* **12** 1–286.

Vranas, K., Jopling, J., Scott, J., Badawi, O., Harhay, M., Slatore, C., Ramsey, M., Breslow, M., Milstein, A. and Kerlin, M. (2018). The association of ICU acuity with outcomes of patients at low risk of dying. *Crit Care Med.* **46** 347–353. MR3468357

Yang, J., Mandal, A. and Majumdar, D. (2016). Optimal designs for 2^k factorial experiments with binary response. *Statistica Sinica* **26** 385–411. MR3726770

Yang, J., Tong, L. and Mandal, A. (2017). D-optimal designs with ordered categorical data. *Statistica Sinica* **27** 1879–1902.