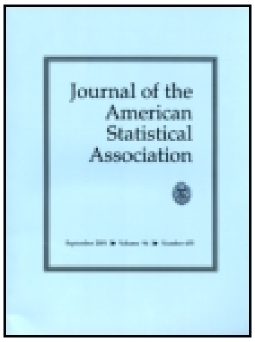

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Minimax efficient random experimental design strategies with application to model-robust design for prediction

Timothy W. Waite & David C. Woods

Minimax efficient random experimental design strategies with application to model-robust design for prediction

Timothy W. Waite*

Department of Mathematics, University of Manchester, UK

and

David C. Woods

Statistical Sciences Research Institute, University of Southampton, UK

*Timothy W. Waite is Lecturer in Statistics, Department of Mathematics, University of Manchester, M13 9PL, United Kingdom (e-mail: timothy.waite@manchester.ac.uk); and David C. Woods is Professor of Statistics, Statistical Sciences Research Institute, University of Southampton, SO17 1BJ, United Kingdom (e-mail: d.woods@southampton.ac.uk).

This work was supported by the UK Engineering and Physical Sciences Research Council through a Doctoral Prize (TWW, provided via Doctoral Training Account EP/K503150/1) and a Fellowship (DCW, EP/J018317/1). Some computations were performed using the Iridis computational facility at the University of Southampton. We are grateful to the reviewers for insightful comments which significantly improved the paper.

Corresponding author Timothy W. Waite timothy.waite@manchester.ac.uk

Abstract

In game theory and statistical decision theory, a random (i.e. mixed) decision strategy often outperforms a deterministic strategy in minimax expected loss. As experimental design can be viewed as a game pitting the Statistician against Nature, the use of a random strategy to choose a design will often be beneficial. However, the topic of minimax-efficient random strategies for design selection is mostly unexplored, with consideration limited to Fisherian randomization of the allocation of a predetermined set of treatments to experimental

units. Here, for the first time, novel and more flexible random design strategies are shown to have better properties than their deterministic counterparts in linear model estimation and prediction, including stronger bounds on both the expectation and survivor function of the loss distribution. Design strategies are considered for three important statistical problems: (i) parameter estimation in linear potential outcomes models, (ii) point prediction from a correct linear model, and (iii) global prediction from a linear model taking into account an L_2-class of possible model discrepancy functions. The new random design strategies proposed for (iii) give a finite bound on the expected loss, a dramatic improvement compared to existing deterministic exact designs for which the expected loss is unbounded.

Key words: Randomization, potential outcomes, statistical decision theory.

1 Introduction

1.1 Decision-theoretic design and random decisions

In frequentist decision-theoretic experimental design, the success of the experiment in relation to its objective is quantified by the value of a loss function, $\ell(\boldsymbol{\theta},\boldsymbol{\alpha})$. Here, the vector $\boldsymbol{\theta}$ contains the true parameter values and is chosen by Nature, while the vector $\boldsymbol{\alpha} = h(\xi, \mathbf{y})$ contains estimates of the interest parameters, $\boldsymbol{\alpha} = a(\boldsymbol{\theta})$. Additionally, $\xi = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denotes the design, to be chosen by the Statistician, and $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ is the vector of responses, yet to be observed.

Prior to the experiment the loss is a random variable and so cannot simply be minimized. Thus the design is instead chosen to give a favourable pre-experimental distribution of possible losses. Usually in the optimal design literature only deterministic choices of ξ are considered. However, we will argue that if Nature is a passive participant rather than a reactive and antagonistic opponent, then one can often obtain a loss distribution with better properties by instead using a suitable random strategy to select ξ.

In the conventional approach, favourability of the loss distribution associated with design ξ is measured by considering the expected loss, $R(\boldsymbol{\theta},\xi) = \mathrm{E}[\ell(\boldsymbol{\theta},\boldsymbol{\alpha})]$, also known as the *risk*. If the risk is independent of $\boldsymbol{\theta}$, then ξ is simply chosen to minimize $R(\boldsymbol{\theta},\xi)$. Otherwise a minimax design is typically used instead, i.e. a ξ_{mM} that minimizes $\Psi(\xi) = \max_{\boldsymbol{\theta}' \in \Theta} R(\boldsymbol{\theta}',\xi)$ with respect to ξ, where Θ is the set of possible values for $\boldsymbol{\theta}$. Most standard 'alphabetic' design optimality criteria for linear models, such as *A*-, *L*- and *V*-optimality, can be derived from this framework via an appropriate choice of loss function. Note that $\Psi(\xi)$ is a tight upper bound for the (unknown) expected loss, $R(\boldsymbol{\theta},\xi)$, attained at the true parameter values. Hence, among all deterministic designs, ξ_{mM} gives the strongest possible bound on the attained expected loss.

Though the above property appears to give a strong argument in favour of the use of ξ_{mM}, in fact in both game theory and statistical decision theory it is widely recognized that a minimax deterministic decision is often outperformed by a randomized decision strategy (e.g. Blackwell and Girshick 1979, Berger 1985, Ch.5, Thie and Keough 2011, Ch.9). Since design selection can be viewed as a decision problem, or alternatively a game pitting the Statistician against Nature, it stands to reason that random decision strategies should also be beneficial for experimental design. Nonetheless, aside from a few minimax analyses of Fisherian randomization (Wu 1981, Li 1983, Hooper 1989, Bhaumik and Mathew 1995), the topic of minimax random strategies for design selection appears almost totally unexplored in the literature.

In this paper, we address this deficiency by introducing the notion of a general random design strategy (RDS) and investigating the performance of a minimax RDS in a number of important design problems. We find that a minimax RDS typically gives a substantially reduced upper bound on the attained expected loss, and a substantially reduced upper bound on the survivor function, i.e. the probability that the loss exceeds a given threshold. It is able to do so by exploiting our key assumption that Nature is passive, and does not change θ in response to our choice for ξ.

The generality of our decision-theoretic formulation enables a wide variety of design problems to be addressed in a unified framework. Specifically, in this paper we consider:

(i) design for parameter estimation in a linear potential outcomes model with fixed (i.e. non-random) unknown unit effects (cf. Bailey 1981, Wu 1981, Dasgupta et al. 2015, Ding 2017).

(ii) design for prediction at an unknown point in a correctly-specified, normal-response linear model.

and, of particular importance, the problem that motivated this work:

(iii) model-robust design for global prediction in a normal-response linear model contaminated by an unknown model discrepancy function belonging to an L_2-class (cf. Wiens 2015).

1.2 Organization of the paper

In Section 2, the definition of a minimax random design strategy is presented, and general bounds are introduced for the survivor function of the loss distribution. The connections between the proposed framework and other related notions in the literature are discussed.

In Section 3, random design strategies are explored for a linear potential outcomes model in which unit-treatment additivity may not hold. We show for the first time that a well-known design and analysis strategy from the additive case remains minimax under these weaker assumptions. In particular, we show that under appropriate conditions the minimax combination of random design strategy and estimator is to use complete randomization of a classical optimal design and ordinary least squares (cf. Li 1983).

Section 4 illustrates for the first time how random design strategies improve minimax efficiency under a loss function extending *G*-optimality, under the assumptions of mean-zero normal unit effects and finite design space. Examples are presented in which a minimax RDS gives a reduced bound on the expected loss and the survivor function of the loss distribution compared to both a deterministic *G*-optimal exact design (cf. St John and Draper 1975) and a *G*-optimal approximate design.

The most interesting results are given in Section 5, where random translation design strategies are proposed for model-robust prediction, in the presence of model discrepancy from an L_2-class. Our new strategies give designs with finitely many runs and bounded expected loss. This represents a major improvement on existing model-robust design theory, in which it is only possible to achieve bounded expected loss by using a design with infinitely many runs, a practical impossibility.

Section 6 contains further discussion of the context of our results. Proofs are deferred to the supplementary material.

2 Random design strategies

2.1 Some terminology and assumptions

The vector $\mathbf{x}_i \in \mathcal{X}, i = 1, \ldots, n$, denotes the treatment applied to the *i*th experimental unit. The set, $\mathcal{X}$, of possible treatments is assumed to be a compact subset of $\mathbb{R}^q$ for some $q \in \mathbb{N}$. Let Ξ denote the set of competing designs, so that $\boldsymbol{\xi} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \Xi$. Throughout this paper, it is assumed that $\Xi = \mathcal{X}^n$; that is, any run order of any choice of *n* treatments from $\mathcal{X}$, not necessarily distinct, is permitted. Clearly this would not be the case if, for example, a multi-stratum design structure were required (Bingham 2015). Given $\boldsymbol{\xi} = \boldsymbol{\xi}'$ and $\boldsymbol{\theta} = \boldsymbol{\theta}'$, the response vector **y** is assumed to be a random draw from the probability measure $P(\cdot \mid \boldsymbol{\xi}', \boldsymbol{\theta}')$ on $\mathbb{R}^n$.

2.2 Random design strategies

Definition 2.1. *A random design strategy (RDS) π is a probability measure on Ξ such that the n-point design ξ to be run in the experiment is chosen at random by sampling from π.*

A design ξ chosen via application of a random design strategy is called a *random design*. Note that a deterministic design is a special case of a random design; it corresponds to a strategy which assigns probability 1 to a particular $\xi' \in \Xi$, i.e. a strategy π with singleton support $\mathrm{supp}(\pi) = \{\xi'\}$. Hence there is no disadvantage to considering random designs; if a random design gives no improvement over a deterministic design, then the optimal π will be a point-mass distribution, $\delta^{\xi'}$. Where emphasis is required, we refer to a RDS with more than one support design as *non-deterministic*. For mathematical precision, we suppose that $\pi(A)$ is defined for any Borel-measurable subset $A \subseteq \Xi$.

The following assumption that Nature is passive and not antagonistic seems plausible in most situations, and is key to our analysis. For a detailed discussion of why this assumption is needed, see Section 6.

Assumption 2.2. $\boldsymbol{\theta}$ *is fixed, i.e. it is chosen independently of* ξ *.*

If Assumption 2.2 holds then, prior to the sampling of a specific design realization, the pre-experimental loss distribution is that induced on $\ell(\boldsymbol{\theta},\boldsymbol{\alpha})$ by the joint distribution of ξ and **y**. The attained pre-experimental expected loss is thus $R(\boldsymbol{\theta},\pi) = \mathrm{E}[\ell(\boldsymbol{\theta},\boldsymbol{\alpha})] = \int_{\Xi}\int_{\mathbb{R}^n} \ell[\boldsymbol{\theta}, h(\xi,\mathbf{y})]dP(\mathbf{y} \mid \xi,\boldsymbol{\theta})d\pi(\xi) = \int_{\Xi} R(\boldsymbol{\theta},\xi)d\pi(\xi)$. This can be given a tight upper bound via $R(\boldsymbol{\theta},\pi) \le \Psi(\pi) = \max_{\boldsymbol{\theta}' \in \Theta} R(\boldsymbol{\theta}',\pi)$.

Definition 2.3. *A minimax RDS,* π_{mM} *, minimizes the upper bound* $\Psi(\pi)$ *with respect to π.*

Note that writing $R(\boldsymbol{\theta},\pi)$ and $\Psi(\pi)$ for the risk and risk bound of an RDS π is a slight abuse of notation as we have already defined $R(\boldsymbol{\theta},\xi)$ and $\Psi(\xi)$ as the risk and risk bound of a deterministic design ξ. However, the concepts are analogous, and it should be clear from the context which is intended.

2.3 Survivor function of the loss distribution

So far we have only considered the expected loss. To analyse the loss distribution in more detail, one can consider its survivor function, $\Pr[\ell(\boldsymbol{\theta},\boldsymbol{\alpha})>u]$ (for a related idea see the quantile criterion of Kapelner et al. 2020). For an RDS π, this is $S(\boldsymbol{\theta},\pi,u)=\int_{\Xi}\int_{\mathbb{R}^n} I\{\ell[\boldsymbol{\theta},h(\boldsymbol{\xi},\mathbf{y})]>u\}dP(\mathbf{y}\mid\boldsymbol{\xi},\boldsymbol{\theta})d\pi(\xi)$, where $I(\cdot)$ denotes an indicator function. For a deterministic design ξ, it is $S(\boldsymbol{\theta},\xi,u)=S(\boldsymbol{\theta},\delta^{\xi},u)=\int_{\mathbb{R}^n} I\{\ell[\boldsymbol{\theta},h(\xi,\mathbf{y})]>u\}dP(\mathbf{y}\mid\xi,\boldsymbol{\theta})$.

The attained survivor function $S(\boldsymbol{\theta},\pi,u)$ is unknown due to its dependence on $\boldsymbol{\theta}$. However it can be bounded tightly in a similar way to the attained expected loss, namely $S(\boldsymbol{\theta},\pi,u)\le\max_{\boldsymbol{\theta}'\in\Theta}S(\boldsymbol{\theta}',\pi,u)$. A weaker bound can be obtained using Markov's inequality, giving $S(\boldsymbol{\theta},\pi,u)\le\min\left\{1,\frac{\Psi(\pi)}{u}I[u\le\ell_{\max}]\right\}$, where $\ell_{\max}=\sup_{(\boldsymbol{\theta}',\xi',\mathbf{y}')}\ell[\boldsymbol{\theta}',h(\boldsymbol{\xi}',\mathbf{y}')]$. This latter bound is not tight, but it can be useful for comparison of a minimax RDS with a minimax deterministic design in cases where $\max_{\boldsymbol{\theta}'\in\Theta}S(\boldsymbol{\theta}',\pi_{\mathrm{mM}},u)$ is difficult to compute. Compared to a minimax deterministic design, a minimax RDS typically gives a reduced upper bound on the survivor function. Figures 1, 2, and 5 illustrate this phenomenon.

2.4 Relationship with other approaches

Fisherian randomization

In practice, an element of randomness in design selection is already commonly recommended by most statisticians. Specifically, it is almost unanimously accepted as beneficial to perform Fisherian randomization of the allocation of treatments to experimental units. This has many advantages, but we focus on two.

First, randomization can be used as a basis for statistical inference without strong modelling assumptions, for example using Neymannian randomization-based estimation or Fisher's exact test. These techniques can only be applied if an appropriate randomization has been used, or alternatively a valid rerandomization procedure (e.g. Morgan and Rubin 2012). They are not applicable with a fully deterministic optimal design (Li et al. 2018).

Second, Fisherian randomization improves experiment robustness, although this benefit is not captured by design performance metrics such as *D*-efficiency. In our view, this inability to explain the advantage of randomization is a regrettable weakness of standard optimal design theory. In contrast, with minimax theory randomization arises as a necessary consequence of optimality under appropriate conditions. For example, in a small series of pioneering papers it was shown that, under uncertainty about the mean of the random errors in a linear model, Fisherian randomization of a standard optimal design is minimax (see Wu 1981, Li 1983, Hooper 1989, Bhaumik and Mathew 1995), deepening the mathematical foundation of longstanding statistical practice.

Despite its many advantages, Fisherian randomization is quite restrictive when viewed within the wider space of general random strategies. As a consequence it will not be optimal in all situations. From the existing literature it is currently unclear what is a minimax strategy when there is uncertainty about aspects of the problem other than the unit effects, for example the functional form of the regression model, or the choice of a location for prediction. Our more flexible approach enables good statistical properties to be obtained in a wider range of problems.

Approximate designs

In addition to the exact designs discussed in Section 1.1, another traditional optimal design approach is to work with an approximate design $\eta = \{\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_K; w_1, \ldots, w_K\}$, with support points $\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_K \in \mathcal{X}$ and weights

$w_1, \ldots, w_K > 0$ $(\sum_{k=1}^{K} w_k = 1)$ (e.g. Kiefer and Wolfowitz 1959). Many numerical methods exist for constructing an approximate design that is optimal with respect to some criterion such as *G*- or *D*-optimality (e.g. Yang et al. 2013, Harman et al. 2020).

Practical implementation of an approximate design depends on the use of a rounding method to determine an integer number, $n_k \approx nw_k$, of runs to be allocated to treatment $\check{\mathbf{x}}_k$, subject to $\sum_{k=1}^{K} n_k = n$. Most commonly the rounding is determined via an optimization procedure. For example, with Kiefer rounding the *n*$_k$ are selected to minimize $\max_k |n_k / n - w_k|$, and with Adams rounding the *n*$_k$ are selected to maximize $\min_k n_k / (nw_k)$ (Pukelsheim 2006, Ch.12). Adams rounding gives an optimal efficiency bound. Another common approach is Federov's method (Pronzato and Pázman 2013, 296-7). We refer to a deterministic (exact) design obtained by one of these procedures as a deterministic ROAD (Rounded Optimal Approximate Design). The choice could also be randomized, by selecting one member uniformly at random from the set of discretizations satisfying the Kiefer or Adams criteria (see Section 4.1). Sometimes this randomized rounding procedure gives a minimax RDS (e.g. Sections 4.1.2 and 4.1.3), other times it does not (e.g. Section 4.1.4).

Note that independent random sampling of points from the approximate design measure exhibits poor properties and is never recommended. For small *n* it gives a non-negligible probability of obtaining a singular exact design. Its asymptotic performance is also inferior compared to other rounding methods. Specifically, under independent sampling the difference between the proportion, n_k / n, of runs allocated to $\check{\mathbf{x}}_k$ and the optimal proportion, *w*$_k$, is of order $O_p(n^{-1/2})$ as $n \to \infty$. In contrast, with the Adams method the difference is of the smaller order $O(1/n)$.

An RDS π is a measure on exact designs, unlike η which is a measure on treatments. Consequently π contains much more detailed information about the experimental procedure than η. For example, unlike η, the RDS π implicitly specifies the probability distribution of the unit-treatment allocation, and any correlations between the replication numbers of two treatments over different realisations of the design. In Section 4 we show that this additional detail can enable superior statistical performance, especially in multifactor experiments with small sample size.

Random balance designs

Despite the superficially similar nomenclature, the approach presented in this paper has little in common with the much-criticised ‘random balance’ designs (Satterthwaite 1959). Our perspective is that it is not the randomness of such designs that is an issue per se, but the poor structure of that randomness, as those strategies are chosen without any decision-theoretic justification. For polynomial response surface models, random balance designs can often lead to problems such as highly correlated parameter estimators or even non-estimability due to partial or total confounding of some factor effects. In contrast, we show that the random strategies proposed in Sections 4 and 5 for the estimation of polynomial response surface models give demonstrably better efficiency than deterministic designs.

3 Randomization in linear potential outcomes models

3.1 Potential outcomes models

The potential outcomes framework was first introduced by Neyman for designed experiments (Splawa-Neyman et al. 1990), and has been extended to observational studies (e.g. Rubin 2005) and factorial experiments (Dasgupta et al. 2015). We denote by $Y_i(\mathbf{x})$ $(i=1,\ldots,n)$ the potential outcome for the response that would occur if the ith experimental unit were to receive treatment $\mathbf{x}\in\mathcal{X}$. The totality of these counterfactual potential outcomes is referred to as the *science*. However, the 'fundamental problem of causal inference' is that only one treatment can be applied per unit and hence the science can only ever be partially observed.

Our assumptions can be more clearly stated after rewriting the potential outcomes as

$$Y_i(\mathbf{x}) = \mu(\mathbf{x}) + \epsilon_i(\mathbf{x}), \qquad (1)$$

where $\mu(\mathbf{x}) = \bar{Y}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{x})$ denotes the mean response under conditions $\mathbf{x}$, and $\epsilon_i(\mathbf{x})$ denotes the *unit effect* of the *i*th unit under treatment $\mathbf{x}$. By construction, the unit effects satisfy $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$. Unlike conventional statistical modelling, in the Neymanian approach the unit effects are treated as fixed unknowns instead of random variables. The only manner in which randomness arises in the observed responses is therefore from the random assignment of treatments to experimental units. There is no need to assume normality or independence of the unit effects. It is even possible to relax the assumption of unit-treatment additivity, equivalent to the condition that $\epsilon_i(\mathbf{x}) \equiv e_i$ for all $\mathbf{x} \in \mathcal{X}$, which is commonly made in experimental design (e.g. Kempthorne 1955, Bailey 1981, 2017).

In the remainder of this section we adopt a *linear* potential outcomes model, in which

$$\mu(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}, \qquad (2)$$

where $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}), \ldots, f_p(\mathbf{x})]^{\mathrm{T}}$, with $f_j : \mathcal{X} \to \mathbb{R}$ ($j = 0, \ldots, p$) a known regression basis function that is continuous with respect to the topology on $\mathcal{X}$, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^{p+1}$ a vector of unknown parameters. We typically assume that $f_0(\mathbf{x}) = 1$, i.e. the model has an intercept. The response actually observed for the *i*th unit, which receives treatment $\mathbf{x}_i$, is

$y_i = Y_i(\mathbf{x}_i) = \mathbf{f}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta} + \epsilon_i(\mathbf{x}_i)$. (3)

In the absence of blocks, the traditional approach to designing an experiment is to use a completely randomized design (CRD). This consists of selecting a deterministic *n*-tuple of treatments, $\boldsymbol{\xi} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and allocating these to experimental units according to a random permutation $\rho \sim \mathrm{Uniform}(S_n)$, giving $\mathbf{x}_i = \mathbf{x}_{\rho^{-1}(i)}$. Here S_n denotes the symmetric group of order *n*.

The probability measure corresponding to the CRD strategy can be denoted concisely using the concept of a pushforward measure. Specifically, the permutation $\rho \in S_n$ acts as a bijection on $\Xi = \mathcal{X}^n$ via $\rho(\boldsymbol{\xi}) = (\mathbf{x}_{\rho^{-1}(1)}, \ldots, \mathbf{x}_{\rho^{-1}(n)})$ for $\boldsymbol{\xi} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \Xi$, and on random design strategies via the pushforward operation $\rho : \pi \mapsto \rho_* \pi$, with $\rho_* \pi(A) = \pi(\rho^{-1}(A))$, with *A* a Borel-measurable subset of Ξ. The probability measure corresponding to the CRD defined above can then be written as $\pi^{\mathrm{CRD},\xi} = \frac{1}{n!} \sum_{\rho \in S_n} \rho_* \delta^{\xi}$, where δ^{ξ} denotes the point-mass probability distribution with support $\{\xi\}$.

Under a CRD, a reordered version of the response vector can be shown to follow a correlated heteroscedastic non-normal linear model, as we see below. Let $r_i = Y_{\rho(i)}(\mathbf{x}_i)$ denote the response from the unit which is allocated to treatment $\mathbf{x}_i$ under the CRD. Further let $\boldsymbol{\epsilon}(\mathbf{x}) = (\epsilon_1(\mathbf{x}), \ldots, \epsilon_n(\mathbf{x}))^{\mathrm{T}}$ denote the vector of unit effects functions and $\mathbf{E}(\boldsymbol{\xi})$ denote the *n* × *n* matrix with *i*, *j*th entry $\epsilon_i(\mathbf{x}_j)$. Also let $S^2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i(\mathbf{x})^2$ denote the variance of the unit effects for treatment $\mathbf{x} \in \mathcal{X}$, $\mathbf{F}_{\xi} = [\mathbf{f}(\mathbf{x}_1) \ldots \mathbf{f}(\mathbf{x}_n)]^{\mathrm{T}}$ denote the model matrix, and $\mathrm{diag}[a_1, \ldots, a_n]$ denote an *n* × *n* diagonal matrix with (*i*, *i*)th entry a_i.

Proposition 3.1. *With the strategy* $\pi^{\mathrm{CRD},\xi}$ *and model (3) the re-ordered response vector* $\mathbf{r} = (r_1, \ldots, r_n)^{\mathrm{T}}$ *satisfies* $\mathrm{E}(\mathbf{r}) = \mathbf{F}_{\xi}\boldsymbol{\beta}, \mathrm{Var}(r_i) = S^2(\mathbf{x}_i)$*, and*

$$\mathrm{Var}(\mathbf{r}) = \mathbf{V}(\boldsymbol{\xi}) = \frac{n}{n-1} \mathrm{diag}[S^2(\mathbf{x}_1), \ldots, S^2(\mathbf{x}_n)] - \frac{1}{n(n-1)} \mathbf{E}(\boldsymbol{\xi})^{\mathrm{T}} \mathbf{E}(\boldsymbol{\xi}) .$$

3.2 Minimax estimator and design strategy

Before identifying the minimax estimator and design strategy, we must first specify the loss function and its corresponding risk, together with our assumptions about the set of possible unit effects and the estimator used.

It is supposed that the experimental goal is estimation of a transformation of the parameters, $\boldsymbol{\alpha} = \boldsymbol{\Lambda\beta}$, where $\boldsymbol{\Lambda}$ is an $r \times (p+1)$ matrix and β_0 is not of interest, so that the first column of $\boldsymbol{\Lambda}$ consists of zeroes. A corresponding loss function is

$$\ell(\boldsymbol{\theta},\boldsymbol{\alpha}) = (\boldsymbol{\alpha}-\boldsymbol{\alpha})^{\mathrm{T}}(\boldsymbol{\alpha}-\boldsymbol{\alpha}), \qquad (4)$$

where $\boldsymbol{\alpha} = h(\boldsymbol{\xi},\mathbf{y})$ is the estimator, not necessarily ordinary least squares, and $\boldsymbol{\theta} = (\boldsymbol{\epsilon},\boldsymbol{\beta}) \in \Theta = \mathcal{E} \times \mathbb{R}^{p+1}$, where $\mathcal{E}$ denotes the set containing all unit effects function vectors $\boldsymbol{\epsilon}$ considered possible prior to the experiment. When computing the risk no integral is needed with respect to **y**, since **y** is uniquely determined given $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$ under model (3). Hence $R(\boldsymbol{\theta},\pi) = \mathrm{E}[\ell(\boldsymbol{\theta},\boldsymbol{\alpha})] = \int_{\Xi} \ell[\boldsymbol{\theta}, h(\xi,\mathbf{y})]d\pi(\boldsymbol{\xi})$ and $\Psi(\pi) = \max_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta},\pi)$.

Two possibilities are considered for the set of possible unit effects, denoted $\mathcal{E}_1$ and $\mathcal{E}_2$, giving rise to two possibilities for Θ, namely $\Theta_1 = \mathcal{E}_1 \times \mathbb{R}^{p+1}$ and $\Theta_2 = \mathcal{E}_2 \times \mathbb{R}^{p+1}$. The first is

$$\mathcal{E}_1 = \Big\{\boldsymbol{\epsilon}(\cdot) = (\epsilon_1(\cdot),\ldots,\epsilon_n(\cdot)) \Big| \epsilon_i : \mathcal{X} \to \mathbb{R} \text{ is measurable,}$$
$$\text{with } \sum_{i=1}^{n} \epsilon_i(\mathbf{x}) = 0 \text{ and } \sum_{i=1}^{n} \epsilon_i(\mathbf{x})^2 \le n\sigma^2 \text{ for all } \mathbf{x} \in \mathcal{X}\Big\}.$$

This set consists of all unit effects function vectors such that $S^2(\mathbf{x}) \le \sigma^2$. This corresponds to the situation where the variances of the potential outcomes are bounded and allowed to differ among treatments, but there is no prior knowledge that a particular treatment has a smaller variance. The second possibility is

$$\mathcal{E}_2 = \{\boldsymbol{\epsilon}(\cdot) \,|\, \boldsymbol{\epsilon}(\mathbf{x}) \equiv \mathbf{e}, \mathbf{e} \in \mathbb{R}^n, \mathbf{e}^{\mathrm{T}}\mathbf{1} = 0, \mathbf{e}^{\mathrm{T}}\mathbf{e} \le n\sigma^2\}$$

which corresponds to the case of unit-treatment additivity with $S^2(\mathbf{x}) \equiv s^2 \le \sigma^2$. The sets $\mathcal{E}_1$ and $\mathcal{E}_2$ are both invariant to permutations of the unit labels, corresponding to an assumption that prior knowledge about the units is homogeneous. If it is instead believed that there is some structure to the experimental units, such as blocking, then a more heterogeneous set of possible unit effects function vectors should be considered.

Several conditions are imposed on the estimator $\boldsymbol{\alpha}$. First, we suppose it is linear, i.e. $\boldsymbol{\alpha} = h(\boldsymbol{\xi}, \mathbf{y}) = \mathbf{A}_{\xi}\mathbf{y}$, as is conventional for Neymanian point estimation (cf. Dasgupta et al. 2015, Zhao et al. 2018). In addition we will suppose that $\boldsymbol{\alpha}$ is *invariant*, i.e. $h(\rho(\xi), \rho(\mathbf{y})) = h(\xi, \mathbf{y})$ for all $\rho \in S_n$. In other words, permutation of the order in which the data are written yields identical estimates. Further, we suppose that $\boldsymbol{\alpha}$ is continuous in the sense that the map $\boldsymbol{\xi}' \mapsto \mathbf{A}_{\xi'}$ is continuous on the support, $\mathrm{supp}(\pi)$, of the RDS. These properties are all satisfied by the ordinary least squares estimator, $\boldsymbol{\alpha}_{\mathrm{OLS}} = \boldsymbol{\Lambda}\mathbf{M}_{\xi}^{-1}\mathbf{F}_{\xi}^{\mathrm{T}}\mathbf{y}$, provided the RDS is *non-singular*, i.e. $\det \mathbf{M}_{\xi'} > 0$ for any $\boldsymbol{\xi}' \in \mathrm{supp}(\pi)$, where $\mathbf{M}_{\xi} = \mathbf{F}_{\xi}^{\mathrm{T}}\mathbf{F}_{\xi}$ denotes the information matrix (see Lemma B.1 in the supplementary material).

Under more restrictive assumptions about the unit effects than adopted here, the optimal combination of estimator and design is well-known. In particular, with normal random errors $\epsilon_i \sim N(0, \sigma^2)$ it is known by the Gauss-Markov theorem that $\boldsymbol{\alpha}_{\mathrm{OLS}}$ is the best linear unbiased estimator. Moreover, in this case the design strategy minimizing the expectation of the loss (4) would be an *L*-optimal deterministic design, i.e. a $\boldsymbol{\xi}_L^* \in \Xi$ that minimizes $\mathrm{tr}[\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{M}_{\xi}^{-1}]$ with respect to $\boldsymbol{\xi} \in \Xi$.

The result below identifies the minimax combination of RDS and estimator for the potential outcomes model (3).

Theorem 3.2. *For model (3) and loss function (4) with* $\Theta = \Theta_1$ *or* Θ_2 *:*

(i) if $\boldsymbol{\alpha}$ is invariant, given any non-singular π the strategy $\tilde{\pi} = \frac{1}{n!}\sum_{\rho \in S_n} \rho_* \pi$ obtained by uniform random permutation of the treatments sampled from π satisfies $\Psi(\tilde{\pi}) \leq \Psi(\pi)$;

(ii) for the strategy $\tilde{\pi}$, the ordinary least squares estimator $\boldsymbol{\alpha}_{\mathrm{OLS}}$ is unbiased and minimax among all continuous linear invariant estimators. With this design strategy and estimator, we have $\Psi(\tilde{\pi}) = \frac{n\sigma^2}{n-1}\int_{\Xi} \mathrm{tr}[\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{M}_{\xi}^{-1}]d\pi(\boldsymbol{\xi})$;

(iii) subject to the constraint that $\boldsymbol{\alpha}$ is continuous, linear and invariant, a minimax combination of RDS and estimator is complete randomization of an L-optimal deterministic design together with $\boldsymbol{\alpha} = \boldsymbol{\alpha}_{\mathrm{OLS}}$.

For a deterministic design, the minimax estimator and maximum risk under the potential outcomes model (3) are given by the following result.

Proposition 3.3. *For model (3), loss function (4), and a deterministic design, ξ, with $\mathbf{M}_{\xi}$ invertible: (i) $\boldsymbol{\alpha}_{\mathrm{OLS}}$ is minimax among all estimators of $\boldsymbol{\alpha}$ for $\Theta = \Theta_1$ or Θ_2; (ii) $\Psi(\pi;\Theta_1) = \max_{\boldsymbol{\theta}' \in \Theta_1} R(\boldsymbol{\theta}', \boldsymbol{\xi}) \geq \max_{\boldsymbol{\theta}' \in \Theta_2} R(\boldsymbol{\theta}', \boldsymbol{\xi}) = \Psi(\boldsymbol{\xi};\Theta_2) = n\sigma^2 \lambda_{\max}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{M}_{\xi}^{-1})$, where $\lambda_{\max}(\cdot)$ denotes the maximal eigenvalue of a matrix; and (iii) the survivor function of the loss distribution satisfies $\max_{\boldsymbol{\theta}' \in \Theta_1} S(\boldsymbol{\theta}', \xi, u) \geq \max_{\boldsymbol{\theta}' \in \Theta_2} S(\boldsymbol{\theta}', \xi, u) = I[u < \Psi(\xi;\Theta_2)]$.*

The max-risk efficiency of RDS π' relative to π is defined as $\mathrm{eff}\,(\pi';\pi) = \Psi(\pi)/\Psi(\pi')$. From Proposition 3.3(ii) we see that, relative to the completely randomized version of ξ, the max-risk efficiency of the unrandomized version of ξ satisfies

$$\mathrm{eff}\,[\delta^{\xi};\pi^{\mathrm{CRD},\xi}] \leq \frac{\mathrm{tr}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{M}_{\xi}^{-1})}{(n-1)\lambda_{\max}(\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Lambda}\mathbf{M}_{\xi}^{-1})} \qquad (5)$$

for $\Theta = \Theta_1, \Theta_2$, with equality when $\Theta = \Theta_2$.

3.2 Example: full quadratic model, three factors, $\mathcal{X} = \{-1, 0, 1\}^3$, *n* = 20 runs

Here $\mathbf{f}(\mathbf{x}) = (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2)^{\mathrm{T}}$ for $\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}} \in \mathcal{X}$ and we set $\mathbf{\Lambda} = [\mathbf{0}_{p\times 1} \,|\, \mathbf{I}_p]$. By Theorem 3.2 the minimax random design strategy, π_{mM}, is to apply Fisherian randomization of the run order to an *L*-optimal design, ξ_L^*. Such a design has been computed using co-ordinate exchange and is given in the supplementary material (Table 2). The risk bound for this strategy is $\Psi(\pi_{\mathrm{mM}}) = 1.34\sigma^2$. Randomization provides a substantial efficiency gain: by (5) the max-risk efficiency of the unrandomized version of ξ_L^* is at most 24.9% relative to π_{mM}.

Figure 1 shows, for each of ξ_L^* and π_{mM}, an upper bound on the attained survivor function of the loss distribution in the case $\Theta = \Theta_2$, using Proposition 3.3(iii) and the Markov bound from Section 2.3. Here we have exploited the fact that the bounds only depend on the ratio u / σ^2 to produce a plot without assuming a specific value of σ^2. We see that, compared to the unrandomized *L*-optimal design, the minimax RDS reduces the worst-case probability of a loss exceeding *u* for all values of *u* shown. Note that by 'reduces', we mean 'gives a value which is less than or equal to the original value'.

From the above, it is clear that randomization provides a substantial benefit. However, it is noteworthy that it is nonetheless inadequate to randomize the run order of a poor set of treatments. For example, the CRD based on ξ_{bad} in Table 2 has a max-risk efficiency of at most 0.9% relative to π_{mM}, far lower than the unrandomized *L*-optimal design.

4 Point prediction in normal-response linear models

4.1 *G*-optimal random design strategies

In this section it is assumed that $\mathcal{X}$ is finite, and that under treatment $\mathbf{x} \in \mathcal{X}$ the response is distributed as $N[\mu(\mathbf{x}), \sigma^2]$ with $\sigma^2 > 0$ unknown. In addition the model is assumed to be linear, i.e. the mean response function satisfies (2).

A classic deterministic optimal design for prediction for this model is the *G*-optimal exact design, $\boldsymbol{\xi}_G^* \in \Xi = \mathcal{X}^n$, which minimizes $\max_{\mathbf{x}' \in \mathcal{X}} \mathrm{Var}\hat{\mu}(\mathbf{x}') = \sigma^2 \max_{\mathbf{x}' \in \mathcal{X}} \mathbf{f}^{\mathrm{T}}(\mathbf{x}')\mathbf{M}_{\xi}^{-1}\mathbf{f}(\mathbf{x}')$ with respect to $\boldsymbol{\xi}$. Above, $\hat{\mu}(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_{\mathrm{OLS}}$ denotes the ordinary least squares prediction of the mean response.

The *G*-optimal exact design above may be derived as a minimax deterministic decision-theoretic design via the choice of an appropriate loss function. To do so, we must assume that the goal of the experiment is to predict $\alpha = \mu(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$ at a point $\mathbf{x} \in \mathcal{X}$, which (a) is not known by the Statistician at the time of planning the experiment, (b) does not change as a result of the choice of ξ, and (c) is known during analysis. For example, this may be the case if the prediction is made by someone else. Then an appropriate loss function is given by the predictive squared error,

$$\ell(\boldsymbol{\theta}, \hat{\alpha}) = [\hat{\alpha} - \alpha]^2, \qquad (6)$$

depending on the unknown $\boldsymbol{\theta} = (\mathbf{x}, \sigma^2, \boldsymbol{\beta}) \in \Theta = \mathcal{X} \times [\underline{\sigma}^2, \overline{\sigma}^2] \times \mathbb{R}^{p+1}$, with $\underline{\sigma}^2$ and $\overline{\sigma}^2$ defining lower and upper bounds on σ^2, respectively. Under an alternative experimental goal, a different loss may be more suitable. For example, if the aim is instead to globally estimate the whole function μ, then integrated squared prediction error may be appropriate. In Section 5, the latter loss function is applied for such global prediction problems in the context of approximate linear models.

To verify that the minimax deterministic design under loss (6) coincides with a *G*-optimal exact design as claimed, first note that it is minimax to set $\hat{\alpha} = \hat{\mu}(\mathbf{x})$ (see Proposition B.5 in the supplementary material). In this case the loss simplifies as

$$\ell(\boldsymbol{\theta}, \hat{\alpha}) = [\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})]^2, \qquad (7)$$

and the risk of a deterministic design $\xi \in \Xi$ becomes

$$R(\boldsymbol{\theta},\boldsymbol{\xi}) = \mathrm{E}_{\mathbf{y}|\boldsymbol{\xi},\boldsymbol{\beta},\sigma^2}\,\ell(\boldsymbol{\theta},\hat{\alpha}) = \mathrm{Var}[\hat{\mu}(\mathbf{x})|\boldsymbol{\xi}] = \sigma^2\tilde{R}(\mathbf{x},\boldsymbol{\xi}), \quad (8)$$

with $\tilde{R}(\mathbf{x},\xi) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\mathbf{M}_{\xi}^{-1}\mathbf{f}(\mathbf{x})$. It is clear from this that the *G*-optimality criterion is equivalent to minimization of $\max_{\boldsymbol{\theta}'\in\Theta} R(\boldsymbol{\theta}',\xi)$ with respect to ξ.

The assumption that **x** does not change in response to the choice of ξ, i.e. assumption (b) above, is analogous to Assumption 2.2. Therefore decision-theoretic arguments indicate that a minimax deterministic design may be outperformed by an RDS *π*. The risk of *π* is $R(\boldsymbol{\theta},\pi) = \mathrm{E}_{\xi,\mathbf{y}|\boldsymbol{\beta},\sigma^2}\,\ell(\boldsymbol{\theta},\hat{\alpha}) = \sigma^2\tilde{R}(\mathbf{x},\pi)$, with $\tilde{R}(\mathbf{x},\pi) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\mathrm{E}\{\mathbf{M}_{\xi}^{-1}\}\mathbf{f}(\mathbf{x})$, and a minimax RDS minimizes $\Psi(\pi) = \bar{\sigma}^2 \max_{\mathbf{x}'\in\mathcal{X}} \tilde{R}(\mathbf{x}',\pi)$ with respect to *π*. We also refer to a minimax RDS as a *G-optimal random design strategy*. If the numbers of elements of $\mathcal{X}$ and Ξ are sufficiently small, then it is possible to obtain a minimax RDS numerically by solving an appropriate linear programming problem, using a standard method from game theory (see Section A in the supplementary material). However, this is not suitable as a general-purpose approach if *n* is large or if $\mathcal{X}$ has a large number of points.

Before proceeding we note that for any non-singular RDS the worst-case survivor function of the loss distribution can be computed using the following result.

Proposition 4.1. *For a normal response linear model* $\mathbf{y}|\xi,\boldsymbol{\beta},\sigma^2 \sim N[\mathbf{F}_{\xi}\boldsymbol{\beta},\sigma^2\mathbf{I}_n]$*, loss function (6), and a non-singular RDS π, the survivor function of the loss distribution satisfies*

$$\max_{\boldsymbol{\theta}'\in\Theta} S(\boldsymbol{\theta}',\pi,u) = \max_{\mathbf{x}'\in\mathcal{X}}\left[\sum_{\xi\in\mathrm{supp}(\pi)}\left(2-2\Phi\left[\frac{\sqrt{u}}{\bar{\sigma}}\{\mathbf{f}^{\mathrm{T}}(\mathbf{x}')\mathbf{M}_{\xi}^{-1}\mathbf{f}(\mathbf{x}')\}^{-1/2}\right]\right)\pi(\boldsymbol{\xi})\right].$$

4.1.1 Approximate designs

A *G*-optimal approximate design η^* for this problem minimizes $\max_{\mathbf{x}\in\mathcal{X}} \mathbf{f}(\mathbf{x})^{\mathrm{T}}\mathbf{M}_{\eta}^{-1}\mathbf{f}(\mathbf{x})$ with respect to the approximate design *η*, where $\mathbf{M}_{\eta} = \int_{\mathcal{X}} \mathbf{f}(\mathbf{x})\mathbf{f}^{\mathrm{T}}(\mathbf{x})d\eta(\mathbf{x})$ is the information matrix. From the general equivalence theorem, this is equivalent to a *D*-optimal approximate design. The following results show that for certain sample sizes a deterministic rounding of η^* can be minimax or highly minimax efficient within the set of RDS, in which case it is not necessary to use a non-deterministic strategy.

Proposition 4.2. *Suppose that* η^* *is minimally supported and n is divisible by p + 1. Then (i) the support points of* η^* *are equally weighted; (ii) the ROAD has* $n/(p+1)$ *replicates of each support point of* η^**; and (iii) the deterministic ROAD is minimax within the set of RDS.*

Proposition 4.3. *The max-risk efficiency of a deterministic ROAD,* ξ_A*, obtained via Adams apportionment of the G-optimal approximate design* η^* *is at least* $1-K/n$*, where K denotes the number of support points of* η^**.*

As an example, consider the implications of these results for a one-factor polynomial regression of degree *d*. It is well-known that the *G*- (equivalently *D*-) optimal approximate design is minimally supported. Hence the deterministic ROAD ξ_A is minimax if *n* is divisible by the number of parameters, or highly efficient if *n* is large. However for small sample sizes or multi-factor problems, a rounding of the *G*-optimal approximate design may be inefficient compared to a minimax RDS. Inefficiency can arise whether the rounding is chosen deterministically (e.g. Sections 4.1.2–4.1.4) or at random from the set of optimal discretizations (e.g. Section 4.1.4).

4.1.2 Example: first-order model, 1 factor, *n* = 3 runs, $\mathcal{X} = \{-1, 0, 1\}$

Here $\mathbf{f}(x) = (1, x)^{\mathrm{T}}$ for $x \in \mathcal{X}$. The minimax random design strategy π_{mM}, found using linear programming, is to choose between $\xi_1 = (-1, -1, 1)$ and $\xi_2 = (-1, 1, 1)$ each with probability $\frac{1}{2}$. This strategy has $\Psi(\pi_{\mathrm{mM}}) = 0.75\bar{\sigma}^2$.

In contrast, the deterministic designs obtained using standard methods are suboptimal. For example, the *G*-optimal deterministic exact design, $\xi_{\mathrm{mM}} = (-1, 0, 1)$

, has a max-risk efficiency of $100\Psi(\pi_{\mathrm{mM}})/\Psi(\xi_{\mathrm{mM}})\% = 90\%$. The *G*-optimal approximate design η^* has two support points, $x = \pm 1$, each with weight $\frac{1}{2}$. Rounding η^* gives either ξ_1 or ξ_2; each of these roundings satisfies both the Kiefer and Adams rounding criteria. Note therefore that in this case the minimax RDS can be viewed as a random choice of one of the two possible Adams roundings of the *G*-optimal approximate design. The randomization is essential: viewed as deterministic exact designs, both ξ_1 and ξ_2 are suboptimal, each with a max-risk efficiency of 75%.

4.1.3 Example: quadratic model, 1 factor, *n* = 4 runs, $\mathcal{X} = \{-1, 0, 1\}$

Here $\mathbf{f}(x) = (1, x, x^2)^{\mathrm{T}}$ for $x \in \mathcal{X}$. For *n* = 4, the minimax RDS assigns an equal probability of $\frac{1}{3}$ to each of the designs $\xi_1 = (-1, 0, 1, -1), \xi_2 = (-1, 0, 1, 0)$, and $\xi_3 = (-1, 0, 1, 1)$, each of which is a Kiefer rounding of the *G*-optimal approximate design $\eta^* = \{-1, 0, 1; \frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$. We have that $\Psi(\pi_{\mathrm{mM}}) = 0.8333\bar{\sigma}^2$, and $\Psi(\xi_1) = \Psi(\xi_2) = \Psi(\xi_3) = \bar{\sigma}^2$. This shows that, once more, randomization of the rounding is essential: considered as a deterministic design, each of ξ_1, ξ_2, ξ_3 has a minimax efficiency of just 83%.

4.1.4 Example: full quadratic model, 2 factors, *n* = 6 runs, $\mathcal{X} = \{-1, 0, 1\}^2$

Here $\mathbf{f}(\mathbf{x}) = (1, x_1, x_2, x_1x_2, x_1^2, x_2^2)^{\mathrm{T}}$ for $\mathbf{x} = (x_1, x_2)^{\mathrm{T}} \in \mathcal{X}$. There are 76 possible non-singular deterministic designs modulo permutations of the run order, which under the model assumptions in Section 4 do not affect the risk function and so are irrelevant. (If weaker assumptions were used, such as in Section 3, then the run order would also need to be considered). The minimax deterministic design has $\Psi(\xi_{\mathrm{mM}}) = 2.75\bar{\sigma}^2$.

A minimax RDS, π_{mM}, was obtained using linear programming (see Figure 2, left panel). This has 8 support designs, $\xi_1, \ldots, \xi_8$, with varying probabilities, and $\Psi(\pi_{\mathrm{mM}}) = 1.55\bar{\sigma}^2$; in fact, $\tilde{R}(\mathbf{x}, \pi_{\mathrm{mM}}) = 1.55$ for all $\mathbf{x} \in \mathcal{X}$. Support designs ξ_1, ξ_5, ξ_6

and ξ_7 are minimax deterministic designs; the remaining four are not but they are helpful in reducing the risk bound for the random design strategy.

The minimax RDS again outperforms traditional deterministic designs. The deterministic *G*-optimal exact design has a max-risk efficiency of $1.55/2.75 \times 100\% = 56\%$ relative to the minimax RDS. The *G*-optimal approximate design for this problem is given in Table 1. The 6-run Kiefer roundings of this approximate design contain all 4 corner points and the center point, plus one edge mid-point. The max-risk efficiency of a fixed (i.e. non-randomized) Kiefer rounding is just 56.4% relative to the minimax RDS. Adams rounding is not recommended here: due to the small sample size, it may give a singular design. Randomizing the choice of Kiefer rounding leads to a strategy with a max-risk efficiency of just 73%. Hence randomized rounding of a *G*-optimal approximate design is inadequate in this example, and the more flexible general RDS is necessary.

Figure 2 (right panel) shows that, compared to several other strategies, the minimax RDS gives a reduced upper bound on the attained survivor function of the loss distribution, $S(\boldsymbol{\theta}, \pi, u)$. In particular, for a wide range of values of u it outperforms the minimax deterministic design, and both fixed and randomized Kiefer roundings of the *G*-optimal approximate design.

5 Model-robust strategies for global prediction

5.1 Model-robust design and approximate linear models

A long-standing problem with many traditional ‘alphabetic’ design optimality criteria is their reliance on an assumed model, which must be specified prior to the experiment by the Statistician. The resulting designs are often inefficient if the true data-generating model differs from the one that has been used to compute the optimal design (Box and Draper 1959). This is in part a consequence of the fact that most design optimality criteria are variance-based; more robust designs may be obtained by accounting for the bias that is introduced if the model is incorrect. For example, suppose that the true data-generating model is $y \sim N[\mu(\mathbf{x}), \sigma^2]$, with μ not necessarily linear, and the Statistician's a priori assumed model for design purposes is the linear model $y \sim N[\mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}, \sigma^2]$. When the experimental goal is global prediction, a common choice for the design is the *V*-optimal design, ξ_V^*, which minimizes $\mathrm{tr}(\mathbf{A}\mathbf{M}_\xi^{-1})$, where $\mathbf{M}_\xi = \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}_i)\mathbf{f}^{\mathrm{T}}(\mathbf{x}_i)$ is the information matrix and $\mathbf{A} = \int_{\mathcal{X}} \mathbf{f}^{\mathrm{T}}(\mathbf{x})\mathbf{f}(\mathbf{x})d\lambda(\mathbf{x})$, with λ Lebesgue measure. Equivalently, the *V*-optimal design minimizes the integrated variance of predictions from the assumed linear model.

Variance-based criteria such as *V*-optimality are reasonable if the assumed model is correct, i.e. if $\mu(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$, because in this case the predictions are unbiased. However, when the assumed model is incorrect, the predictions are biased, and this should be accounted for in the design. For example, we might evaluate design performance using the integrated mean squared error of predictions from the linear model. On this basis, Box and Draper (1959) found that a variance-minimizing design is often outperformed by a purely bias-minimizing design. However, their conclusions were limited to the somewhat unrealistic case where the true model μ is a polynomial, and the assumed linear model is a polynomial of lower degree.

A more flexible approach to model-robust design can be achieved by allowing the true and assumed model means to differ by an essentially arbitrary function. More precisely, authors such as Wiens (2015) suppose that

$$\mu(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_{\mathrm{ba}} + \psi(\mathbf{x}), \qquad \text{for some } \psi \in \mathcal{H}, \qquad (9)$$

where ψ is a discrepancy function that represents the error that results from approximating μ with a linear model. The class $\mathcal{H}$ is chosen to include all discrepancy functions considered possible a priori by the Statistician; most commonly it is defined as

$$\mathcal{H} = \left\{ \psi \in \mathcal{L}^2(\mathcal{X};\lambda) \middle| \int_{\mathcal{X}} [\psi(\mathbf{x})]^2 \, d\lambda(\mathbf{x}) \le \tau^2, \int_{\mathcal{X}} \psi(\mathbf{x})\mathbf{f}(\mathbf{x}) \, d\lambda(\mathbf{x}) = \mathbf{0}_{p+1} \right\}, \qquad (10)$$

where $\mathcal{L}^2(\mathcal{X};\lambda)$ denotes the set of real-valued functions on $\mathcal{X}$ that are square-integrable with respect to λ (e.g. Wiens 1992, Heo et al. 2001, Dette and Wiens 2009). Other choices are possible, for example based on a uniform bound or a smoothness class for ψ (Li and Notz 1982, Yue and Hickernell 1999). Note that orthogonality condition in (10) is perfectly natural: it corresponds to the assumption that the parameter values $\boldsymbol{\beta}_{\mathrm{ba}}$ give the best linear model approximation to the true model, as measured by the L_2 distance. To see this, note that the L_2-best approximating parameter values $\boldsymbol{\beta}_{\mathrm{ba}}$ satisfy

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\beta}} \int_{\mathcal{X}} [\mu(\mathbf{x}) - \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}]^2 \, d\lambda(\mathbf{x}) \Bigg|_{\boldsymbol{\beta}_{\mathrm{ba}}} = -\int_{\mathcal{X}} [\mu(\mathbf{x}) - \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_{\mathrm{ba}}]\mathbf{f}(\mathbf{x}) \, d\lambda(\mathbf{x}) = -\int_{\mathcal{X}} \psi(\mathbf{x})\mathbf{f}(\mathbf{x}) \, d\lambda(\mathbf{x}).$$

An appropriate decision-theoretic formulation can be developed by considering the experimental goal of global prediction. In this case the interest 'parameter' is $\boldsymbol{\alpha} = \mu$, which can be estimated via $\hat{\mu}(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}$. A suitable loss function is then the integrated squared prediction error of predictions from the assumed linear model, i.e.

$$\ell(\boldsymbol{\theta}, \boldsymbol{\alpha}) = \int_{\mathcal{X}} [\mu(\mathbf{x}) - \mathbf{f}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}]^2 \, d\lambda(\mathbf{x}), \qquad (11)$$

where $\boldsymbol{\theta} = (\psi, \boldsymbol{\beta}_{\mathrm{ba}}, \sigma^2) \in \Theta = \mathcal{H} \times \mathbb{R}^{p+1} \times [\underline{\sigma}^2, \bar{\sigma}^2]$. The corresponding risk, $R(\boldsymbol{\theta}, \boldsymbol{\xi})$, is the integrated mean squared prediction error. It can be shown that the risk is independent of $\boldsymbol{\beta}_{\mathrm{ba}}$, and so the maximum risk satisfies

$$\sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \boldsymbol{\xi}) = \sup_{(\psi, \sigma^2) \in \mathcal{H} \times [\underline{\sigma}^2, \bar{\sigma}^2]} R(\psi, \sigma^2, \boldsymbol{\xi}).$$

Although the discrepancy class (10) has the advantages of being flexible and well-studied, to date it has been troublesome to use when the treatment space is uncountably infinite, e.g. $\mathcal{X} = [-1,1]^q$. In this case, deterministic designs with finitely many runs have woeful performance: it can be shown that any such $\boldsymbol{\xi}'$

has unbounded expected risk, i.e. $\Psi(\xi') = \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \xi') = \infty$, even if *n* is large (Wiens 1992). Even worse, the survivor function has the undesirable property given in Proposition 5.1 below. Roughly speaking, the loss is almost sure to exceed any finite bound in the worst case, due to the possibility of arbitrarily unfavourable states of Nature.

Proposition 5.1. *For model (9), loss function (11), and a deterministic design,* ξ*, we have that* $\sup_{\boldsymbol{\theta}'\in\Theta} S(\boldsymbol{\theta}', \boldsymbol{\xi}, u) = 1$ for all $u \geq 0$.

To address the poor performance of finite deterministic designs, the existing literature proposes the use of an optimal design with infinitely many support points, defined through a probability density function *f* on $\mathcal{X}$ (e.g. Wiens 2015). However such an *f* is not practically useful. To obtain a feasible experiment, *f* must be approximated by a design ξ with finitely many points, yet if chosen deterministically any such approximation will suffer from the same problems outlined above. Hence nothing is gained by constructing an optimal *f*. Our novel solution to this paradox is to instead use a random translation design strategy (see Section 5.2). As we show, such a strategy leads to an experiment with bounded risk and improved bounds on the survivor function of the loss distribution.

5.2 Random translation design strategies

Here we adopt model (9), L_2-discrepancy class (10), and loss function (11). We assume that the compact design space $\mathcal{X} \subseteq \mathbb{R}^q$, for integer *q* > 0, has Lebesgue measure $\lambda(\mathcal{X}) > 0$.

The risk for an arbitrary non-singular random design strategy, *π*, can be written as a bias-variance decomposition,

$$R(\psi, \sigma^2, \pi) = \text{MIV}(\sigma^2, \pi) + \text{MISB}(\psi, \pi), \qquad (12)$$

where the mean integrated variance (MIV) and mean integrated squared bias (MISB) are given by $\mathrm{MIV}(\sigma^2,\pi)=\sigma^2\mathrm{E}_\xi\mathrm{tr}(\mathbf{AM}_\xi^{-1}),\ \mathrm{MISB}(\psi,\pi)=b(\psi,\pi)+\|\psi\|_2^2$ and $b(\psi,\pi)=\mathrm{E}_\xi\{\boldsymbol{\psi}_\xi^{\top}\mathbf{K}_\xi\boldsymbol{\psi}_\xi\}$. Above, $\mathbf{A}=\int_{\mathcal{X}}\mathbf{f}(\mathbf{x})\mathbf{f}^{\mathrm{T}}(\mathbf{x})d\lambda(\mathbf{x}),\mathbf{M}_\xi=\mathbf{F}_\xi^{\mathrm{T}}\mathbf{F}_\xi$ is the information matrix, $\|\psi\|_2^2=\int_{\mathcal{X}}\psi(\mathbf{x})^2d\lambda(\mathbf{x})$ is the L_2-norm, and $\boldsymbol{\psi}_\xi=[\psi(\mathbf{x}_1)\ldots\psi(\mathbf{x}_n)]^{\mathrm{T}}$ is the vector of evaluations of ψ on design ξ. The *bias-sensitivity* matrix $\mathbf{K}_\xi=\mathbf{F}_\xi\mathbf{M}_\xi^{-1}\mathbf{AM}_\xi^{-1}\mathbf{F}_\xi^{\mathrm{T}}$ quantifies the effect of the discrepancy function on the bias of the predictions.

In the case of no discrepancy, i.e. $\psi(\mathbf{x})=0$ for all **x**, the minimax random design strategy reduces to a point-mass measure on the traditional deterministic *V*-optimal design, ξ_V^*. To see this, note that (12) becomes $R(0,\sigma^2,\pi)=\sigma^2\mathrm{E}_\xi\mathrm{tr}(\mathbf{AM}_\xi^{-1})\geq\sigma^2\mathrm{tr}(\mathbf{AM}_{\xi_V^*}^{-1})=R(0,\sigma^2,\delta^{\xi_V^*})$.

To find minimax efficient random strategies, we need to be able to compute the tight risk bound, $\Psi(\pi)=\sup_{(\psi',\sigma'^2)\in\mathcal{H}\times[\underline{\sigma}^2,\bar{\sigma}^2]}R(\psi',\sigma'^2,\pi)=\bar{\sigma}^2\mathrm{E}_\xi\mathrm{tr}(\mathbf{AM}_\xi^{-1})+\sup_{\psi'\in\mathcal{H}}b(\psi',\pi)+\tau^2.$ One potential approach would be to devise an algorithm to numerically maximize $b(\psi',\pi)$ with respect to the function ψ'. However, such an algorithm would likely be extremely computationally intensive. Thus, it is desirable to obtain analytical formulae for $\Psi(\pi)$. For general design strategies, this remains an open problem for future research. However, we have successfully identified a flexible class of random design strategies for which $\Psi(\pi)$ is analytically tractable.

Definition 5.2. *The strategy π is a non-singular random translation design strategy with mean design* $\bar{\boldsymbol{\xi}}=(\mathbf{c}_1,\ldots,\mathbf{c}_n)\in\Xi$ *and translation set* $\mathcal{T}\subseteq\mathbb{R}^q$*, denoted* $\pi=\pi^{\mathrm{RT}}(\bar{\xi},\mathcal{T})$*, if:*

(i) $\mathcal{T}$ *is convex and compact,* $\mathbf{0}_q\in\mathcal{T}$*, and* $\int_{\mathcal{T}}\mathbf{t}d\lambda(\mathbf{t})=\mathbf{0}_q$*;*

(ii) the sets $\mathbf{c}_i+\mathcal{T}=\{\mathbf{c}_i+\mathbf{t}|\mathbf{t}\in\mathcal{T}\}$ *are subsets of* $\mathcal{X}$ *and are almost disjoint in the sense that* $\lambda[(\mathbf{c}_i+\mathcal{T})\cap(\mathbf{c}_{i'}+\mathcal{T})]=0$ *for* $i\neq i'$*;*

(iii) the continuous function $\mathbf{d}:\mathcal{T}\to\Xi$ *defined by* $\mathbf{d}(\mathbf{t})=(\mathbf{c}_1+\mathbf{t},\ldots,\mathbf{c}_n+\mathbf{t})$ *is such that* $\det\mathbf{M}_{\mathbf{d}(\mathbf{t})}>0$ *for all* $\mathbf{t}\in\mathcal{T}$*, so that π is a non-singular strategy;*

(iv) if ξ is distributed according to π, then $\xi = (\mathbf{x}_1,\ldots,\mathbf{x}_n) = \mathbf{d}(\mathbf{t})$, i.e. $\mathbf{x}_i = \mathbf{c}_i + \mathbf{t}, i = 1,\ldots,n$*, with* $\mathbf{t} \sim \mathrm{Uniform}(\mathcal{T})$.

To interpret this definition, note that if $\mathbf{t} \sim \mathrm{Uniform}(\mathcal{T})$ then $\mathrm{E}(\mathbf{t}) = \int_{\mathcal{T}} \mathbf{t} d\lambda(\mathbf{t}) = \mathbf{0}_q$, and so $\mathbf{c}_i$ describes the mean location of the *i*th design point over all the different potential realizations of the random design. The size of the set $\mathcal{T}$ of possible translations determines the degree of randomness; when $\mathcal{T}$ is small the design is close to deterministic. The regularity condition (ii) in Definition 5.2 states that the support sets for the different design points must be non-overlapping (apart from a set of measure zero) and is needed to prove Theorem 5.3. The realized design $\xi = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$ is a translation of the mean design $\bar{\xi} = (\mathbf{c}_1,\ldots,\mathbf{c}_n)$ by a common vector **t** that is sampled randomly according to a uniform distribution on $\mathcal{T}$. Thus ξ retains the same geometric shape as $\bar{\xi}$ (see Figure 3).

For random translation designs, the risk bound is given by the following theorem.

Theorem 5.3. *For a non-singular random translation design strategy* $\pi = \pi^{\mathrm{RT}}(\bar{\xi},\mathcal{T})$*, with* $\lambda(\mathcal{T}) > 0$*,*

$$\Psi(\pi) = \sup_{\boldsymbol{\theta}'\in\Theta} R(\boldsymbol{\theta}',\pi) = \bar{\sigma}^2 \mathrm{E}_{\xi} \mathrm{tr}(\mathbf{A}\mathbf{M}_{\xi}^{-1}) + \tau^2 + \frac{\tau^2}{\lambda(\mathcal{T})} \max\nolimits_{\mathbf{t}\in\mathcal{T}} \lambda_{\max}[\mathbf{K}_{\mathbf{d}(\mathbf{t})}]. \qquad (13)$$

An obvious choice is to set $\mathcal{T} = [-\frac{\delta}{2},\frac{\delta}{2}]^q$ above, with $\delta \geq 0$. Lemma 5.4 below gives necessary and sufficient conditions for such a choice to satisfy condition (ii) of Definition 5.2 in order to give a valid random translation design strategy. Non-singularity of the strategy must be checked separately. We refer to such a design strategy as a *hypercuboidal random translation design strategy*, denoted $\pi^{\mathrm{H}}(\bar{\xi},\delta)$. Note that $\lambda(\mathcal{T}) = \delta^q$, and so the conditions of Theorem 5.3 hold only for $\delta > 0$. Nonetheless equation (13) remains valid for $\delta = 0$ provided $\tau^2 > 0$. This is true because for $\delta = 0$ the design strategy is deterministic and it is known in this case that $\Psi(\pi) = \infty$ (Wiens 1992).

Lemma 5.4. *Suppose that* $\mathcal{X} = [-1,1]^q$. *The choice* $\bar{\boldsymbol{\xi}} = (\mathbf{c}_1, \ldots, \mathbf{c}_n) \in \mathcal{X}^n$ *with* $\mathcal{T} = [-\frac{\delta}{2}, \frac{\delta}{2}]^q$ *satisfies condition (ii) of Definition 5.2 if and only if: (i)* $-1 + \frac{\delta}{2} \leq c_{ij} \leq 1 - \frac{\delta}{2}$, *for* $i = 1, \ldots, n,\ j = 1, \ldots, q$; *and (ii)* $\min_{i \neq i' \in \{1,\ldots,n\}} \left\| \mathbf{c}_i - \mathbf{c}_{i'} \right\|_\infty \geq \delta$.

5.3 Numerical examples

5.3.1 Numerical optimization of hypercuboidal strategies

In order to numerically optimize the strategy $\pi^{\mathrm{H}}(\bar{\xi}, \delta)$, we first need to approximate the objective function $\Psi(\pi) = \Psi(\bar{\xi}, \delta)$ in (13). We do so via two steps: (i) Monte Carlo estimation of $\mathrm{E}_\xi \mathrm{tr}(\mathbf{A}\mathbf{M}_\xi^{-1})$ using a Latin Hypercube sample, $\mathbf{s}_1, \ldots, \mathbf{s}_K \in [-1,1]^q$, of potential (scaled) translation vectors, and (ii) approximate maximization of $\lambda_{\max}(\mathbf{K}_{\mathbf{d}(\frac{\delta}{2}\tilde{\mathbf{t}})})$ with respect to $\tilde{\mathbf{t}} \in [-1,1]^q$ via a finite discretization, $\mathcal{T} = \{\tilde{\mathbf{t}}_1, \ldots, \tilde{\mathbf{t}}_M\}$, of $[-1,1]^q$. This gives

$$\hat{\Psi}(\bar{\xi}, \delta) = \frac{\bar{\sigma}^2}{K} \sum_{k=1}^{K} \mathrm{tr}(\mathbf{A}\mathbf{M}^{-1}_{\mathbf{d}(\frac{\delta}{2}\mathbf{s}_k)}) + \tau^2 + \frac{\tau^2}{\delta^q} \max_{\tilde{\mathbf{t}} \in \mathcal{T}} \lambda_{\max}(\mathbf{K}_{\mathbf{d}(\frac{\delta}{2}\tilde{\mathbf{t}})}) .$$

Using a reparameterization, we may recast the problem of optimizing $\bar{\xi}$ and δ subject to the complicated constraints in Lemma 5.4 as a simpler box-constrained problem. Specifically, we work in terms of $\xi = (\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_n) = (\mathbf{c}_1 / (1 - \frac{\delta}{2}), \ldots, \mathbf{c}_n / (1 - \frac{\delta}{2}))$ and $\tilde{\delta} = \delta / \min_{i \neq i'} \left\| \mathbf{c}_i - \mathbf{c}_{i'} \right\|_\infty$, noting that $\bar{\xi}, \delta$ satisfy the constraints of Lemma 5.4 if and only if $-1 \leq \tilde{c}_{ij} \leq 1$ and $0 \leq \tilde{\delta} \leq 1$. Hence minimization of $\hat{\Psi}(\bar{\xi}, \delta)$ with respect to $(\bar{\xi}, \delta)$ is equivalent to minimization of $\hat{\Psi}_2(\xi, \tilde{\delta}) = \hat{\Psi}\left[2\xi / (2 + \tilde{\delta} \min_{i \neq i'} \left\| \tilde{\mathbf{c}}_i - \tilde{\mathbf{c}}_{i'} \right\|_\infty), \frac{2\tilde{\delta} \min_{i \neq i'} \left\| \tilde{\mathbf{c}}_i - \tilde{\mathbf{c}}_{i'} \right\|_\infty}{2 + \tilde{\delta} \min_{i \neq i'} \left\| \tilde{\mathbf{c}}_i - \tilde{\mathbf{c}}_{i'} \right\|_\infty} \right]$ with respect to $(\xi, \tilde{\delta})$ subject to the box constraints $\tilde{c}_{ij} \in [-1,1], \tilde{\delta} \in [0,1]$. We address this box-constrained problem using multiple random initializations of a cyclic co-ordinate descent algorithm (cf. Gotwalt et al. 2009). The discretization $\mathcal{T}$ is refined iteratively (cf. Pronzato and Pázman 2013, p.311; for details see the supplementary material).

5.3.2 Simple illustrative examples

For illustration we present approximately minimax hypercuboidal random translation design strategies for the following problems: (i) n = 3 runs, q = 1 factor, and an approximate quadratic model, i.e. $\mathbf{f}(x) = (1, x, x^2)^{\mathrm{T}}$, and (ii) n = 4 runs, q = 2 factors, and an approximate first-order model, i.e. $\mathbf{f}(\mathbf{x}) = (1, x_1, x_2)^{\mathrm{T}}, \mathbf{x} = (x_1, x_2)^{\mathrm{T}}$. Minimax strategies were identified using the approach described in Section 5.3.1 and are plotted in Figure 4 for a range of values of $\tau^2 / \bar{\sigma}^2$. For both problems, it is clear that the minimax choice for $\bar{\xi}$ is similar to the V-optimal deterministic design (for q = 1, $\xi_V^* = (-1, 0, 1)$; for q = 2, ξ_V^* is the 2^2 factorial), modified to account for the constraints of Lemma 5.4. The minimax choice for δ increases as τ^2 increases, i.e. if protection is sought against a discrepancy function with larger L_2 norm, then an RDS with greater variance must be used.

5.3.3 Heuristics for larger examples

In our experience, for problems with larger dimensionality it is computationally expensive to identify a global optimum of $\Psi(\bar{\xi}, \delta)$ using the brute-force optimization approach described in Section 5.3.1. However, the results for the simple illustrative examples suggest that Heuristic 5.5 below may be adequate to identify a combination of $\bar{\xi}, \delta$ with high minimax efficiency. The heuristic essentially performs a one-dimensional optimization to robustify a V-optimal deterministic design. The associated computational cost is minimal, even in problems with a large number of factors and runs. Nonetheless, the gain in robustness compared to the V-optimal design is dramatic, as the resulting RDS has a bounded expected loss.

Heuristic 5.5. *To construct an efficient strategy:*

1. *Calculate a V-optimal or highly V-efficient exact design, ξ_V^*, e.g. by using co-ordinate descent to minimize $\mathrm{tr}(\mathbf{AM}_\xi^{-1})$ with respect to ξ;*

2. *Form a parameterized mean design,* $\bar{\boldsymbol{\xi}}_\delta$*, that approximates* ξ_V^* *and which satisfies the constraints of Lemma 5.4. To do this, move the points of* ξ_V^* *away from the boundary of* $[-1,1]^q$ *if necessary, and split any replicates;*
3. *Choose δ to minimize* $\hat{\Psi}(\bar{\boldsymbol{\xi}}_\delta, \delta)$*.*

In order to facilitate comparison of the heuristic and brute-force approaches, we now give details of an example with a moderate number of factors and runs. Consider an *n* = 12 run design in *q* = 3 factors for an approximately quadratic model, that is $\mathbf{f}(\mathbf{x}) = (1, x_1, x_2, x_3, x_2x_3, x_1x_3, x_1x_2, x_1^2, x_2^2, x_3^2)^{\mathrm{T}}$, $\mathbf{x} = (x_1, x_2, x_3)^{\mathrm{T}}$, with $\mathcal{X} = [-1,1]^3$ and $\tau^2 = 0.02\bar{\sigma}^2$. With around 70 minutes of computation, the brute-force approach identified a strategy $\pi_1 = \pi^{\mathrm{H}}(\bar{\xi}_1, \delta_1)$ with $\Psi(\pi_1) = 6.840\bar{\sigma}^2$. Alternatively, a strategy may be found for this example using Heuristic 5.5. For Step 2 above, we require a $\bar{\xi}_\delta$ that approximates ξ_V^* and satisfies the constraints of Lemma 5.4. This can be obtained from ξ_V^* by setting $c_{ij} = 1 - \frac{\delta}{2}$ (respectively, $c_{ij} = -1 + \frac{\delta}{2}$) when the corresponding element of ξ_V^* is + 1 (respectively, -1), and replacing the replicated points $(-0.092, -0.093, 0.093), (-0.092, -0.093, 0.093)$ with $(-0.092 - \frac{\delta}{2}, -0.093, 0.093), (-0.092 + \frac{\delta}{2}, -0.093, 0.093)$ (for full details see the supplementary material). Figure 5 shows $\hat{\Psi}(\bar{\xi}_\delta, \delta)$ as a function of *δ*. The optimal value of *δ*, used in Step 3 above, is $\delta^* = 0.271$. The resulting heuristic strategy has a risk bound of approximately $7.03\bar{\sigma}^2$, corresponding to a max risk efficiency of 97% relative to π_1. Computation of this efficient heuristic strategy, $\pi_2 = \pi^{\mathrm{H}}(\bar{\xi}_{\delta^*}, \delta^*)$, requires only a few seconds. This is around two orders of magnitude less than the brute-force search.

With no discrepancy (i.e. if $\psi(\mathbf{x}) \equiv 0$) the risk bound from the *V*-optimal deterministic design would be $3.918\bar{\sigma}^2$, compared with $7.03\bar{\sigma}^2$ in the presence of discrepancy if the heuristic random strategy is used. Thus, provided one uses an efficient random strategy, the presence of discrepancy only leads to a 34% increase in the bound on the root mean integrated squared prediction error. In contrast, if a deterministic V-optimal design is used, the presence of discrepancy leads to an unbounded risk.

Figure 5 also shows upper bounds on the survivor function of the loss distribution for both the optimal heuristic strategy π_2 and a deterministic strategy (all deterministic strategies have the same tight bound). For most values of u the random strategy provides a substantially reduced bound on the probability of a loss exceeding u.

Of course, use of a random translation design strategy results in reduced variance-efficiency if in fact the model is correct. We quantify this loss using V-efficiency, $\mathrm{eff}_V(\boldsymbol{\xi}) = R(0,\sigma^2,\boldsymbol{\xi}_V^*) / R(0,\sigma^2,\boldsymbol{\xi}) = \mathrm{tr}\left(\mathbf{AM}_{\xi_V^*}^{-1}\right) / \mathrm{tr}\left(\mathbf{AM}_{\xi}^{-1}\right)$, defined assuming that $\psi(\mathbf{x}) \equiv 0$. Note that $\mathrm{eff}_V(\boldsymbol{\xi})$ is a random variable due to dependence on ξ. Figure 5 shows that for π_2 the realized design typically has a V-efficiency of around 70%. This seems more than adequate given that the random strategy provides such dramatic improvements in robustness.

6 Discussion

We believe that the results in this paper highlight untapped potential for novel random design strategies to lead to substantial improvement in the properties of the loss distribution for a variety of experimental design problems. We anticipate that future research will realise these benefits in diverse areas where there is a priori uncertainty, including design for nonlinear models and screening experiments.

The discussion below focusses on two main themes. First, we clarify assumptions and potential misconceptions, for example the importance of Assumption 2.2 and ideas about optimality over repeated samplings and conditional risk. Second, connections with other areas are explored, including the Bayesian interpretation of randomization and links with the computer model calibration literature.

Importance of Nature’s passivity

We now illustrate the importance of Assumption 2.2 in obtaining improved bounds on the expected loss and the survivor function of the loss distribution. Without it a minimax RDS would give no advantage over a minimax deterministic design. To see this suppose that, instead of θ and ξ being independent, it were possible for θ to be a function of ξ. In this case the attained pre-experimental expected loss of π would not necessarily be given by $R(\boldsymbol{\theta},\pi)=\int_{\Xi}R(\boldsymbol{\theta},\xi)d\pi(\boldsymbol{\xi})$. For example, suppose that after observing our choice of ξ, but before generating **y** from $P(\cdot|\xi,\boldsymbol{\theta})$, Nature chooses a $\boldsymbol{\theta}\in\mathrm{argmax}_{\theta'\in\Theta}R(\boldsymbol{\theta}',\xi)$. In this case, the pre-experimental expected loss under π would be $\int_{\Xi}\max_{\theta'\in\Theta}R(\boldsymbol{\theta}',\xi)d\pi(\xi)\geq\max_{\theta'\in\Theta}R(\boldsymbol{\theta}',\xi_{\mathrm{mM}})$. Hence it would be impossible to improve upon the bound on the expected loss that is given by the minimax deterministic design. However, it seems implausible and unduly pessimistic to suppose that Nature behaves in such a reactive, intelligent and antagonistic manner. In the more realistic case that Assumption 2.2 holds, a minimax deterministic design will be inefficient compared to a minimax RDS due to its focus on guarding against this extreme pathological behaviour. In contrast, a minimax RDS is able to reduce the probability of large losses by exploiting the fact that Nature cannot change $\boldsymbol{\theta}$.

Optimality over repeated samplings

In common with all other optimal frequentist procedures, the minimax RDS is derived using an expectation over hypothetical realizations of the same experiment. This may cause some to be concerned that use of a design sampled from an RDS is optimal only if the same experiment is repeated over and over, when in fact it is only conducted once. However, this concern is unwarranted. Neyman’s original justification for frequentist procedures that minimize expected risk is that if they are applied consistently in many *different* experiments then the total achieved loss across all experiments will be reduced (Berger 1984). Note that this point is not unique to our proposed method. Similar repeated sampling properties are also used when deriving traditional deterministic optimal designs, which typically minimize the variance of a point estimator. This variance is also computed by taking an expectation over hypothetical realizations of the same experiment.

Misconceptions about conditional risk

A related concern is that if one considers only the conditional risk, $R(\boldsymbol{\theta},\xi)$, then at first sight it may appear that there are some drawbacks to the use of a ξ sampled from a minimax RDS rather than a minimax deterministic design. However, these are based on flawed reasoning, and so they should not discourage the use of a minimax RDS.

The first apparent drawback is as follows: once one has chosen ξ, the attained risk could be as large as $\max_{\boldsymbol{\theta}'\in\Theta} R(\boldsymbol{\theta}',\xi)$. The latter is usually larger than the maximum conditional risk $\max_{\boldsymbol{\theta}'\in\Theta} R(\boldsymbol{\theta}',\xi_{\mathrm{mM}})$ that applies if a minimax deterministic design is used. This begs the question: has use of a minimax RDS really reduced the risk? More careful consideration shows that it is indeed very likely to have reduced the risk, because with high probability our random sampling procedure will have generated a ξ with $R(\boldsymbol{\theta},\xi) < \max_{\boldsymbol{\theta}'\in\Theta} R(\boldsymbol{\theta}',\xi_{\mathrm{mM}})$. For example, with the minimax random strategy for linear model prediction in Section 4.1.4, we have that (i) a tight lower bound on the probability is $\Pr[R(\boldsymbol{\theta},\xi) < \max_{\boldsymbol{\theta}'\in\Theta} R(\boldsymbol{\theta}',\xi_{\mathrm{mM}})] \geq 0.684$; and (ii) the probability that $R(\boldsymbol{\theta},\xi) > \max_{\boldsymbol{\theta}'\in\Theta} R(\boldsymbol{\theta}',\xi_{\mathrm{mM}})$ is at most 0.118. (For details of these calculations see the supplementary material).

A second apparent drawback is that, if one considers only the conditional risk, it may seem that use of a minimax RDS has the disadvantage of replacing a certain experimental outcome with an uncertain one. However, this is simply not the case: the realized *loss* is uncertain regardless of whether one uses a deterministically- or randomly-selected design. As shown earlier (e.g. in Figures

1, 2, and 5) a minimax RDS typically gives stronger bounds on the properties of the distribution of possible losses.

Links with Bayesian approach

Here we have focussed on the minimax decision-theoretic framework. From a Bayesian perspective, randomized decision-making is often regarded as unnecessary (e.g. Lindley 1982). However, even in this context randomization has several advantages. First, it simplifies Bayesian causal inference (Rubin 1978). Second, randomization has been shown to be a Bayesian decision-theoretically optimal design strategy in situations where several parties have differing prior information or when the analyst, or final decision-maker, is a different person from the one designing the study (Berry and Kadane 1997, Bonassi et al. 2009). It may be interesting to investigate optimal Bayesian random design strategies in more complex experiments with multiple stakeholders than the simple settings described in the existing literature.

Computer model calibration

We briefly note some similarities and differences between the formulation in Section 5 and calibration of a computer simulator of a physical process (e.g. Kennedy and O'Hagan 2001). In the calibration literature, the basic idea is to approximate the expected response of the physical process under conditions $\mathbf{x}$ using the simulator output, $\eta(\mathbf{x},\boldsymbol{\theta})$. However, before predictions can be made, physically realistic values of the parameters $\boldsymbol{\theta}$ must be determined. This can be done by combining data from physical experiments on the real process with data from a computer experiment on the simulator. A major challenge is that, due the high computational expense of simulator runs, the value of $\eta(\mathbf{x},\boldsymbol{\theta})$ can only be computed for a few combinations of inputs $\mathbf{x},\boldsymbol{\theta}$, necessitating the construction of a computationally cheaper approximation of η, known as an *emulator*.

Similar to our approach in Section 5, in calibration it is assumed that the true mean of the physical process differs from the simulator output by an explicit model discrepancy function, namely $\mathrm{E}(y) = \mu(\mathbf{x}) = \eta(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{ba}}) + \psi(\mathbf{x})$ (cf. (9)). Here $\boldsymbol{\theta}_{\mathrm{ba}}$ is a vector of parameter values giving a best approximation to the physical process. Recently developed L_2-calibration approaches impose an orthogonality condition similar to the one in (10) (Tuo and Wu 2015, Plumlee 2017). An important difference is that in calibration the discrepancy function ψ is explicitly estimated, using Gaussian process techniques, whereas in Section 5 predictions are made without estimating ψ. It would be interesting to explore further connections between model-robust design and calibration in future research.

Supplementary material

The online supplementary material contains proofs of the theoretical results and some additional supporting numerical results. R code for the examples is available from the journal website and via the first author's personal website, at https://github.com/timwaite/random-designs.

References

Bailey, R. A. (1981), 'A unified approach to design of experiments', *J. Roy. Statist. Soc. A* **144**, 214–223.

Bailey, R. A. (2017), 'Inference from randomized (factorial) experiments', *Statist. Sci.* **32**, 352–355.

Berger, J. O. (1984), The frequentist viewpoint and conditioning, *in* L. LeCam and R. Ohlsen, eds, 'Proceedings of the Berkeley Conference in Honor of Kiefer and Neyman', Wadsworth, Belmont.

Berger, J. O. (1985), *Statistical decision theory and Bayesian analysis*, 2nd edn, Springer, New York.

Berry, S. M. and Kadane, J. B. (1997), 'Optimal Bayesian randomization', *J. Roy. Statist. Soc. B* **59**, 813–819.

Bhaumik, D. K. and Mathew, T. (1995), ‘Minimaxity of randomized optimal designs with respect to a general optimality criterion’, *Sankhya B* **57**, 122–127.

Bingham, D. (2015), Multistratum fractional factorial designs, *in* A. Dean, M. Morris, J. Stufken and D. Bingham, eds, ‘Handbook of Design and Analysis of Experiments’, Chapman and Hall/CRC, New York, chapter 8, pp. 321–338.

Blackwell, D. A. and Girshick, M. A. (1979), *Theory of games and statistical decisions*, Dover, New York.

Bonassi, F. V., Nishimura, R. and Stern, R. B. (2009), ‘In defense of randomization: A subjectivist Bayesian approach’, *AIP Conference Proceedings* **1193**, 32–39.

Box, G. E. P. and Draper, N. R. (1959), ‘A basis for the selection of a response surface design’, *J. Amer. Statist. Assoc.* **54**, 622–654.

Dasgupta, T., Pillai, N. S. and Rubin, D. B. (2015), ‘Causal inference from 2^k factorial designs by using potential outcomes’, *J. Roy. Statist. Soc. B* **77**, 727–753.

Dette, H. and Wiens, D. P. (2009), ‘Robust designs for 3D shape analysis with spherical harmonic descriptors’, *Statist. Sinica* **19**, 82–102.

Ding, P. (2017), ‘A paradox from randomization-based causal inference’, *Statist. Sci.* **32**, 331–345.

Gotwalt, C. M., Jones, B. A. and Steinberg, D. M. (2009), ‘Fast computation of designs robust to parameter uncertainty for nonlinear settings’, *Technometrics* **51**, 88–95.

Harman, R., Filová, L. and Richtárik, P. (2020), ‘A randomized exchange algorithm for computing optimal approximate designs of experiments’, *J. Amer. Statist. Assoc.* **115**, 348–361.

Heo, G., Schmuland, B. and Wiens, D. P. (2001), ‘Restricted minimax robust designs for misspecified regression models’, *Canad. J. Statist.* **29**, 117–128.

Hooper, P. M. (1989), ‘Minimaxity of randomized optimal designs’, *Ann. Statist.* **17**, 1315–1324.

Kapelner, A., Krieger, A. M., Sklar, M., Shalit, U. and Azriel, D. (2020), ‘ Harmonizing optimized designs with classic randomization in experiments’, *The American Statistician* pp. 1–12.

Kempthorne, O. (1955), ‘The randomization theory of experimental inference’, *J. Amer. Statist. Assoc.* **50**, 946–967.

Kennedy, M. C. and O’Hagan, A. (2001), ‘Bayesian calibration of computer models’, *J. Roy. Statist. Soc. B* **63**, 425–464.

Kiefer, J. and Wolfowitz, J. (1959), ‘Optimum designs in regression problems’, *Ann. Math. Statist.* **30**, 271–294.

Li, K.-C. (1983), ‘Minimaxity for randomized designs: Some general results’, *Ann. Statist.* **11**, 225–239.

Li, K.-C. and Notz, W. (1982), ‘Robust designs for nearly linear regression’, *J. Statist. Plan. Infer.* **6**, 135–151.

Li, X., Ding, P. and Rubin, D. B. (2018), ‘Asymptotic theory of rerandomization in treatment–control experiments’, *Proceedings of the National Academy of Sciences* **115**(37), 9157–9162.

Lindley, D. V. (1982), ‘The role of randomization in inference’, *Proceedings of the Biennial Meeting of the Philosophy of Science Association* **2**, 431–446.

Morgan, K. L. and Rubin, D. B. (2012), ‘Rerandomization to improve covariate balance in experiments’, *Ann. Statist.* **40**(2), 1263–1282.

Plumlee, M. (2017), ‘Bayesian calibration of inexact computer models’, *J. Amer. Statist. Assoc.* **112**, 1274–1285.

Pronzato, L. and Pázman, A. (2013), ‘Design of experiments in nonlinear models’ , *Lecture notes in Statistics* **212**.

Pukelsheim, F. (2006), *Optimal design of experiments*, SIAM, Philadelphia.

Rubin, D. B. (1978), ‘Bayesian inference for causal effects: the role of randomization’, *Ann. Statist.* **6**, 34–58.

Rubin, D. B. (2005), ‘Causal inference using potential outcomes: Design, modeling, decisions’, *J. Amer. Statist. Assoc.* **100**, 322–331.

Satterthwaite, F. E. (1959), ‘Random balance experimentation’, *Technometrics* **1**, 111–137.

Splawa-Neyman, J., Dabrowska, D. M. and Speed, T. P. (1990), ‘On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.’, *Statist. Sci.* **5**, 465–472.

St John, R. C. and Draper, N. R. (1975), ‘D-optimality for regression designs: a review’, *Technometrics* **17**, 15–23.

Thie, P. R. and Keough, G. E. (2011), *An Introduction to Linear Programming and Game Theory*, Wiley, Hoboken.

Tuo, R. and Wu, C.-F. J. (2015), ‘Efficient calibration for imperfect computer models’, *Ann. Statist.* **43**, 2331–2352.

Wiens, D. P. (1992), ‘Minimax designs for approximately linear regression’, *J. Statist. Plan. Infer.* **31**, 353–371.

Wiens, D. P. (2015), Robustness of design, *in* A. Dean, M. Morris, J. Stufken and D. Bingham, eds, ‘Handbook of Design and Analysis of Experiments’, Chapman and Hall/CRC, New York, chapter 20, pp. 457–470.

Wu, C.-F. (1981), ‘On the robustness and efficiency of some randomized designs ’, *Ann. Statist.* **9**, 1168–1177.

Yang, M., Biedermann, S. and Tang, E. (2013), ‘On optimal designs for nonlinear models: a general and efficient algorithm’, *J. Amer. Statist. Assoc.* **108**, 1411–1420.

Yue, R.-X. and Hickernell, F. J. (1999), ‘Robust designs for fitting linear models with misspecification’, *Statist. Sinica* **9**, 1053–1070.

Zhao, A., Ding, P., Mukerjee, R. and Dasgupta, T. (2018), ‘Randomization-based causal inference from split-plot designs’, *Ann. Statist.* **46**, 1876–1903.

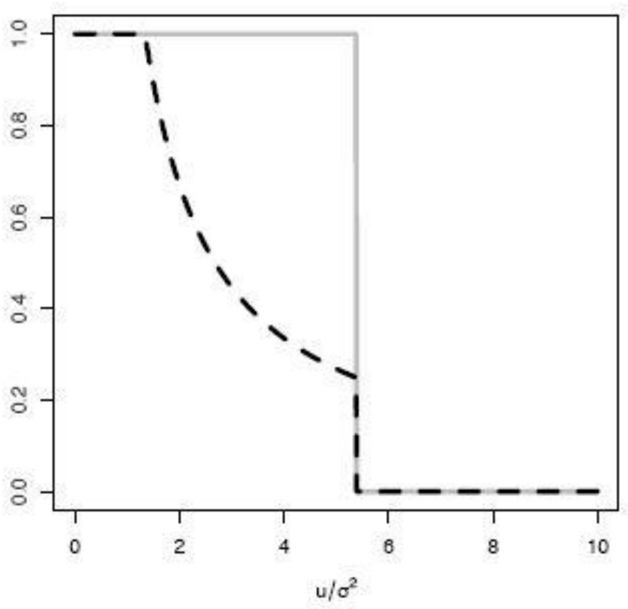

Fig. 1 Upper bounds on the attained survivor function of the loss distribution, $S(\boldsymbol{\theta}, \pi, u)$, for two design strategies in the example of Section 3.2, in the case $\Theta = \Theta_2$. Grey line: unrandomized *L*-optimal design, ξ_{mM} (tight bound). Black line: randomized *L*-optimal design, π_{mM} (Markov bound).

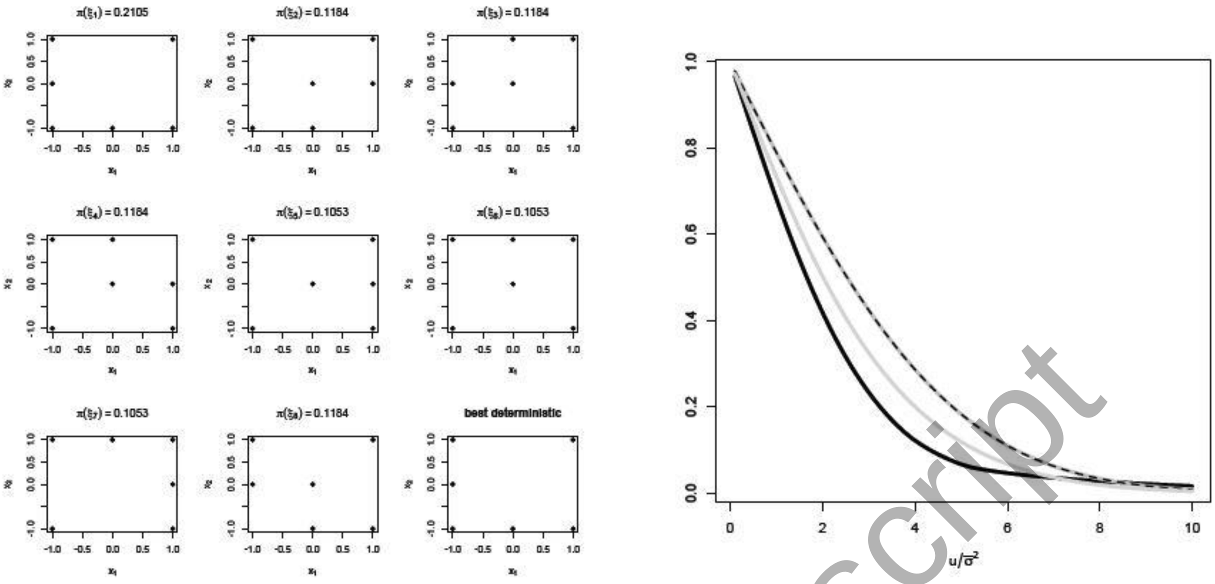

Fig. 2 *Left:* minimax random design strategy for the two factor problem Example 4.1.4: *p* = 5, *n* = 6, $\mathcal{X} = \{-1, 0, 1\}^2$. *Right:* Upper bounds on the attained survivor function of the loss distribution, $S(\boldsymbol{\theta}, \pi, u)$, for four design strategies in the example of Section 4.1.4. Grey dashed line: minimax deterministic design, ξ_{mM} . Black dashed line: fixed Kiefer rounding of the *G*-optimal approximate design. Grey solid line: randomized Kiefer rounding. Black solid line: minimax RDS, π_{mM} .

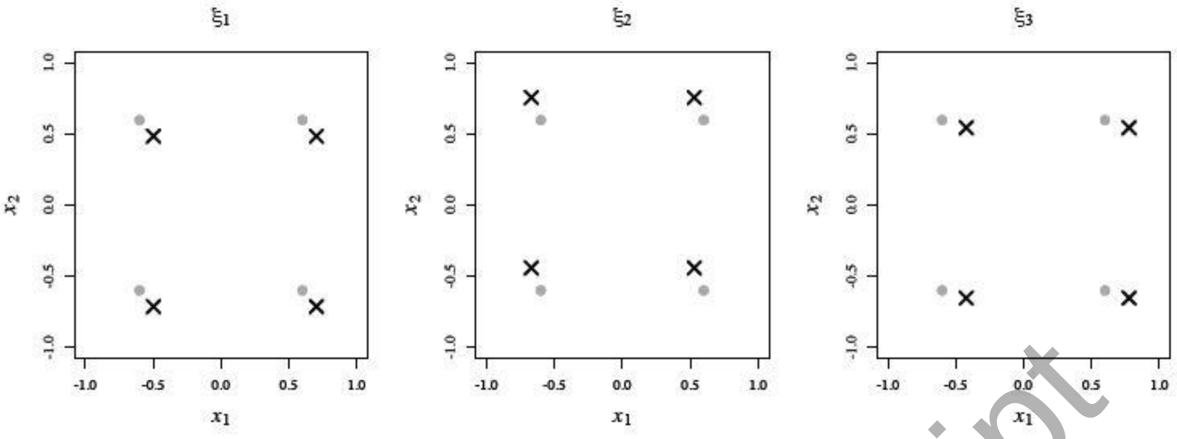

Fig. 3 Three realizations of a random translation design (× – design realization, ξ_i; ● – mean design, $\bar{\xi}$)

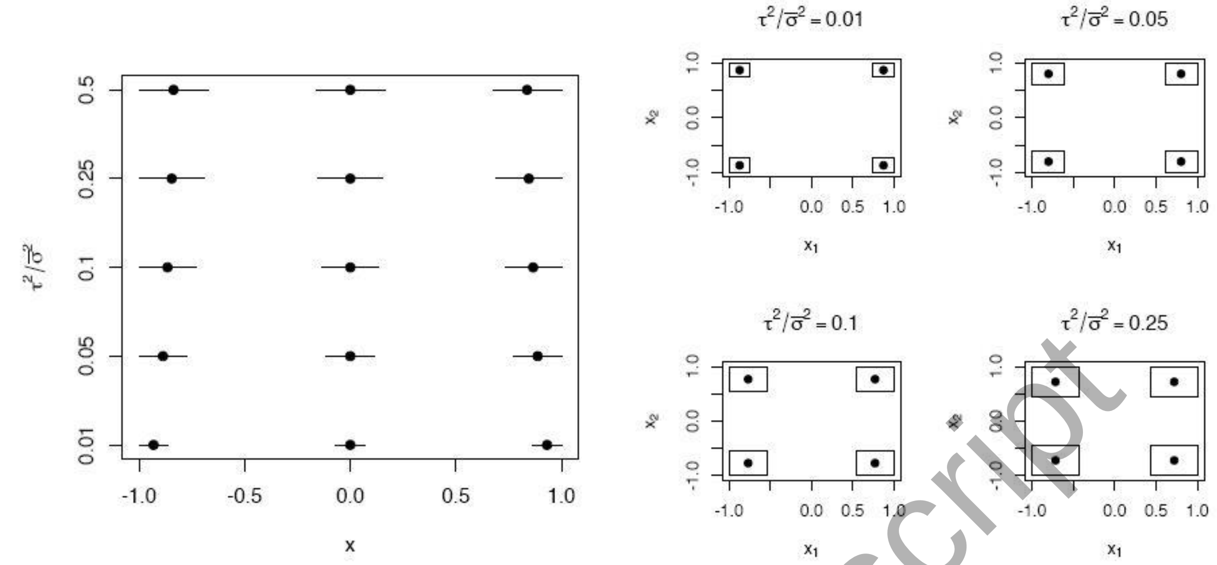

Fig. 4 Minimax random hypercuboidal translation design strategies for the examples in Section 5.3.2, for several values of $\tau^2/\bar{\sigma}^2$. *Left:* approximate quadratic model, *n* = 3 runs, *q* = 1 factor (• indicates $\bar{\xi} = (c_1, c_2, c_3)$, horizontal lines indicate the intervals $c_i \pm \frac{\delta}{2}$). *Right:* approximate linear model, *n* = 4 runs, *q* = 2 factors (• indicates $\bar{\xi} = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4)$, boxes indicate $\mathbf{c}_i + [-\frac{\delta}{2}, \frac{\delta}{2}]^2$).

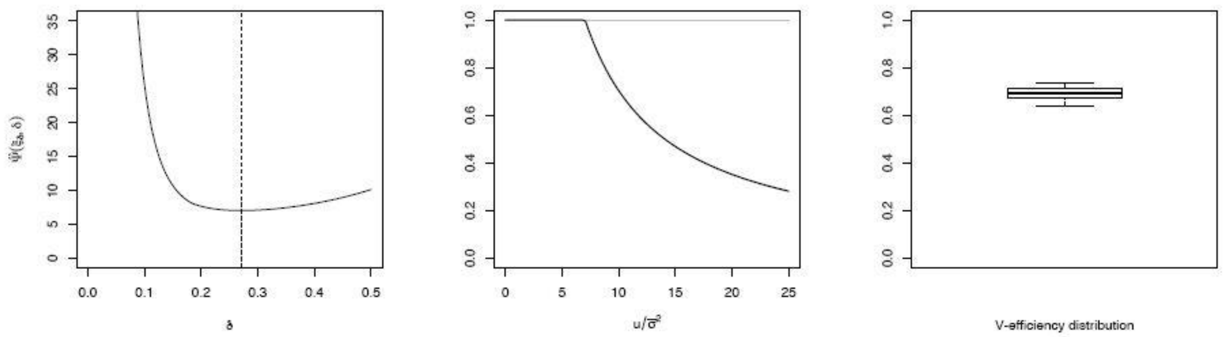

Fig. 5 Example: full quadratic model, *n* = 12, *q* = 3. *Left:* approximate risk bound $\hat{\Psi}(\bar{\xi}_\delta, \delta)$ for the heuristic strategy with different values of δ (vertical line: optimal value, $\delta^* = 0.268$). *Centre:* bounds on the survivor function of the loss distribution, $S(\boldsymbol{\theta}, \pi, u)$ (Black line: random strategy $\pi = \pi^{\mathrm{H}}(\bar{\xi}_{\delta^*}, \delta^*)$, Markov bound. Grey line: any deterministic design, tight bound from Proposition 5.1). *Right: V*-efficiency distribution of $\pi^{\mathrm{H}}(\bar{\xi}_{\delta^*}, \delta^*)$.

Table 1 *G*-optimal approximate design for Example 4.1.4

x_1	–1	0	1	–1	0	1	–1	0	1
x_2	–1	–1	–1	0	0	0	1	1	1
weight	0.1458	0.0802	0.1458	0.0802	0.0962	0.0802	0.1458	0.0802	0.1458