Noname manuscript No.
(will be inserted by the editor)

Optimal replicates for Designed Experiments under the on-line framework

Nandan Sudarsanam · **Balaji Pitchai Kannu** · **Daniel D. Frey**

Received: date / Accepted: date

Abstract This paper explores the use of designed experiments in an on-line environment. Motivated by real-world examples, we model a scenario where the practitioner is given a finite set of units and needs to select a subset of these which are expended towards a one-shot, multi-factor designed experiment. Following this phase, the designer is left with the remaining set of unused units to implement any learnings from the experiments. With this setting, we answer the key design question of how much to experiment, which translates to choosing the number of replicates for a given design. We construct a Bayesian framework that captures the expected cumulative gain across the entire set of units. We derive theoretical results for the optimal number of replicates for all two-level, full and fractional factorial designs with 7 factors or fewer. We conduct simulations that serve as validation of the theoretical results, as well as enabling us to explore scenarios and techniques of analysis that are not captured in the theoretical studies. Our overall results indicate that the optimal allocation of units for experimentation varies from 1% to 20% of the total units available, which is mainly governed by the experimental environment and the total number of units. We conclude that experimenting with the optimal number of replicates recommended by our study can lead to a cumulative improvement which is $80-95\%$ greater than the expected cumulative improvement gained when a practitioner chooses the number of replicates randomly.

Keywords On-line Experimentation · Design of Experiments · Bayesian Analysis

Nandan Sudarsanam
Assistant Professor, Department of Management Studies, Robert Bosch Center for Data Science and AI,
Indian Institute of Technology Madras; E-mail: nandan@iitm.ac.in

Balaji Pitchai Kannu
PhD Candidate, Department of Management Studies,
Indian Institute of Technology Madras; E-mail: ms15d004@smail.iitm.ac.in

Daniel D. Frey
Professor, Mechanical Engineering,
Massachusetts Institute of Technology; E-mail: danfrey@mit.edu

1 Introduction

There has been a vigorous exchange of views within the community of Research in Engineering Design regarding the relationship of design theory to concepts of rationality (Reich (2010)). Much of the debate has centered on challenges of maintaining consistency of our decisions with declared principles and stated preferences. An alternative viewpoint is that correspondence is a more fitting criterion by which to judge decision making methods. Katsikopoulos (2009) argued that a decision process can be viewed as good when its performance in its authentic environment is good even if the process lacks internal logical coherence and therefore could, in theory, lead to poor decisions in some cases. In a position paper that summarizes this debate, Reich (2010) ends by asking researchers to explore several different topics including new issues/criteria that must be addressed when dealing with selection methods. That topic is the one taken up in the present paper although admittedly we constrain our attention to a narrow slice of selection methods, viz, gathering information in a live setting in a way that optimally supports selection.

An important development relevant to this debate about decision making has been the explosion of new theory and applications in data analytics and machine learning. We appear to be at the beginning of a decades-long trend toward increasingly data-intensive, evidence-based decision making (Jordan and Mitchell, 2015). This is highly relevant to design theory because, in many cases, machine learning algorithms engage in selection among alternatives. Notably, the approach that has dominated ML so far is strongly correspondence-based. The best current algorithms do not result in self-consistent rule sets for rational decision making but rather they are pattern recognition schemes that explicitly define success as providing good results across large empirical training sets.

In the age of Machine Learning, the data have proven to be the carriers of value more so than the data processing methods. The most successful companies make little effort to protect algorithms (in fact they tend to distribute them freely) and, by contrast, are aggressive in gathering and protecting data sets. It's worth reflecting what lessons may be garnered for design theory. Perhaps we will find that good design decisions will increasingly be the ones made based on the best data sets (once we understand what best really means). While best data sets are sometimes the largest ones, in other settings they might be the highest quality data sets with the most relevant and timely information. Experimentation is an approach intended to generate such useful data.

Emerging from the rapid expansion of machine learning has been a vibrant academic community studying on-line experimentation through the *bandit* framework (Sutton and Barto, 1998). In the multi-armed bandit problem, a limited set of resources must be allocated between competing (alternative) choices in a way that maximizes their expected gain. This is essential when the algorithms for machine learning are not simply trained once on a predefined data set but instead have to continuously learn in an on-line setting. This raises the important question of how much resource to spend on exploration (which fosters more robust learning) and exploitation (which provides benefits from having learned).

In the context of design, our decisions are specific enough that they require a data gathering effort highly tailored to the task at hand. Our study presents a common real-world context that spans multiple domains and entails the use of experimentation for improvement. In such a setting, the key task for the statistician is to determine an appropriate experimental plan along with the required replicates. This, is often done with the experimental budget in mind. The important trade-off here is to balance the cost of experimentation with the cost of failing to accurately characterize the effect of a particular input factor. The latter could lead to the exclusion of an input factor from the model, or in some cases, it could result in an inference favoring a less than optimal setting. With this dichotomy in place, academics and practitioners have come up with various approaches to determine the number of replicates for a given design (we discuss these in greater detail in Section 2).

In this general form, such a conception assumes that the experiments are constructed in an offline environment. Here, the cost of experimentation is quantified through various channels such as the cost of downtime, labor costs, material costs, operational cost, etc. The product or process resulting from the experiment however does not reach the end user. This might not be feasible in a wide range of environments where the experimentation is conducted on the real system–as opposed to a model–and downtime is not an option. In such systems, the costs of experimenting are, in a sense, the costs associated with producing inferior products. This is precisely the setting that is studied in bandit framework in reinforcement learning. However, the bandit conception of the problem is one of sequential learning. This might be very useful for wide range of applications, such as online Ads, recommender systems, where bandit algorithms have been successfully deployed. There exist a wide range of applications where a planned multi-variable experimental design needs to be constructed, before interacting with the system. In particular, we contend that experiments in production-related setups and design initiatives, owing to the time lines for experimental implementation and delayed feedback of the response, would require a parallel one-shot approach that is typified in DoE. We illustrate this through two concrete examples:

1. Example A (Production): A foundry has an order to create 1000 Aluminum casts. They would like to create casts with minimum hydrogen gas porosity (all casts have some porosity and are still fit for use but the quality of our work is measured by the porosity). They can make minor adjustments to three furnace settings, where each can be set at two levels (eight treatment combinations). They would like to experiment with the furnace parameters to determine which settings minimize porosity for this particular part. The eight furnace settings can be used to separately create a pre-specified volume of molten metal to be poured into a predetermined number of casts (for each of the eight combinations).[1] Using the results from the one-shot experiment, the remaining (1000 minus the number of casts already made) are set to the best settings. How many casts should we commit to experimenting on the furnace settings?

[1] In this case, it would be impractical to operate the furnace, adaptively, by melting metal and changing settings for one cast at a time as the bandit framework would require

2. Example B (Design): A boutique athletic footwear brand decides to put out a set of limited edition shoes, three months before the Rio Olympics 2016. They anticipate a maximum demand of 10,000 pairs and have sourced the materials required for this. The brand is considering two different color themes, and two different sizes of the logo on the shoes. Their objective is to minimize the unsold units before the Olympics. The manager decides to run an experiment by trying each of the four possible combinations. She will pick a winner based on 30 days of sales data from the experiment to commit the remaining units (10,000 minus the shoes that were used in the experiment). How many of the 10,000 shoes does she commit to the experiments? Again, in this example more adaptive approaches–like the bandit framework–are precluded by the fact that the reward is delayed (it takes 30 days of sales data to make meaningful conclusions) and the decision-making can support at most one experimental iteration.

This is the environment we choose to study. One, where a one-shot design like an A/B test, RCT, or a designed experiment is required to improve a system or product. However, the experiments are conducted in an on-line setting where the resulting outcomes are not discarded. Additionally, a constraint is imposed on the total number of units which can be subject to the treatments, since an infinite potential to exploit the findings of an experiment would logically justify infinite experimentation.

For a given experimental plan, the critical design question then becomes to choose the number of replicates that the statistician needs to conduct in this on-line environment. This is the overarching question that this study chooses to address. Intuitively, one can see that there is a trade-off. In example A, if the experimenter chooses to conduct a 100 replicates across the eight treatments, she might find the optimal one with sufficient statistical certainty. However, this would leave her with only 200 casts to implement her knowledge. Whereas, if the experimenter ran a single replicate, she might be exposed to a higher risk of making a sub-optimal conclusion on the factor-levels for the remaining 992 units.

In this study, we present the mathematical framework to determine the optimal number of replicates under certain assumptions on the environment. We specifically consider the case of factorial designs. We demonstrate our approach by deriving the optimal replicates across the entire range of 2-level full factorial and fractional factorial designs for 7 factors or fewer. We augment this study with a computational analysis that looks at different approaches of analyzing designed experiments as well as accounting for various regularities in the environment.

The rest of this paper is structured as follows: In section 2 we present the literature pertinent to this study. In section 3 we present the main results of the paper for all the 2-level full and fractional factorial designs with 7 factors or fewer. In section 4, we present an empirical validation of the theory, and also provide indicative numerical results for the different environments through different techniques. Finally, we conclude our work in the section 5 and suggest some areas for future work.

2 Related Work

Given the multidisciplinary nature of our study, the areas of related literature are also manifold. First, we discuss related work from the perspective of the design philosophy we adopt. This is broadly aligned with other studies that look at the effect of interventions over time rather than a single point optimization. We then study the body of literature pertaining to the use of statistical methods in design engineering to improve processes and products. Finally, we discuss the more methodologically relevant literature; we go beyond the design engineering context and study literature related to the statistics inspired question of sample size determination, in both off-line and on-line scenarios.

From a design philosophy perspective, this paper builds upon a tradition of design scholarship in which the behavior of companies and teams are assessed using a long-term perspective. Decisions can be evaluated not only on the basis of the coherence of the process they use Katsikopoulos (2009) or the probabilistic properties of outcomes from a single project, but rather can include consideration of the future positioning of the company on many later projects. Whitney (1993) studied the product design and risk management strategies of Nippondenso Co. Ltd by conducting extensive on-site visits and interviews over a period from 1974 to 1991. Whitney observed that a key to Nippondensos success was a framing of their decisions by classifying products according to the demands that they placed on their manufacturing infrastructure. In addition, developments were made in manufacturing infrastructure to enable families and sequences of products to be delivered as the needs of their partner firms evolved. As part of their Jikigata-Ken approach, action plans commit resources to improvement efforts that, in some ways, parallel the commitment to a period of experimentation that we seek to guide in this paper. Following in that vein of design scholarship, Gonzalez-Zugasti et al. (2000) considered the ways that product platforms can be architected to enable high performance and flexibility. They studied the behavior of Team X at JPL as a family of interplanetary missions were designed. They documented an iterative process in which platforms were chosen, portions of the design were frozen, and subsets of variables were altered to improve performance. The relationship to the present study arises because the challenges of space missions drive an emphasis on flexibility and performance over a long time horizon. Although Team X did not necessarily field a sequence of missions during the scope of the study (Gonzalez-Zugasti et al., 2000), they did document a design process adaptable to such a sequential improvement process extending over the launch and design of several missions.

This paper, similar to many others in the design literature, builds from a basis of statistical theory and technique. Osio and Amon (1996) developed an adaptive engineering design methodology based on sequential sampling. Similar to the current paper, they support the designer in information gathering for the purpose of performance improvement. Their method was not for the on-line setting and instead used computational models of the design and surrogates thereof rather than physically instantiated systems in the hands of the intended users. Their procedure was Bayesian relying on a prior judgment of the designers to establish a stochastic process reflecting their initial level of confidence which was then updated by data from computer

simulations. Their process of sampling the space was based on A optimal and D optimal designs with no replication (as is common in computer experiments) whereas the current paper employs full and fractional factorial design with replication since the on-line setting will exhibit variation across replicates. Martin and Ishii (1997) used empirical data from industry to guide development of quantitative tools. They developed indices for commonality, differentiation, and set-up and proposed to combine them in a regression based decision procedure. Orsborn et al. (2008) employed Principal Components Analysis to determine the characteristics within vehicle classes that could then be used by product designers to form new designs incorporating derived shape relationships. İç (2016) developed an optimization method combining statistical DoE with the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). This paper is similar to the current one in that it applies DoE to gain information optimally under resource constraints but it differs in its focus on multi-criterion decision problems whereas we assume that a single criterion has been formed to guide decision making.

Independent of the design engineering context, there are multiple well established approaches in the statistics literature which focus on determining the sample size for an experiment. These broadly rely on frequentist inferencing. These foundations are commonly used in single factor experiments which are extended to the multi-factor settings. Montgomery (2008) summarizes three broad approaches to determining sample size: (i) The operational characteristic (OC) curve, (ii) Maximum permissible standard deviation increase, and (iii) Confidence interval estimation methods. All approaches essentially require the experimenter to specify probabilistic requirements on the ability of the experiment to determine differences in the treatments. In multi factor settings the detection of differences also accounts for interactions between treatments of different factors. This concept has also led to simplified heuristics, such as those put forth by Berndtson (1991); Schmidt and Launsby (1989). Other extensions include explorations to multiple response types (Machin et al., 2011; Newcombe, 1998), and to the use of phases of experiments to account for lack of knowledge of critical parameters (Simon, 1989; Simon et al., 1985; Stein, 1945). While these approaches help with the traditional experimental design setup, they are less useful in the on-line setting. There are two key differences. The first is that the probability of failing to attain statistical significance when there is a known difference in treatment mean, also referred to as the Type II error, is not relevant to us. The practitioner is required to choose a treatment combination irrespective of statistical significance. Therefore we adopt a framework that uses a similar approach to the OC curve, but instead quantify the probability of erroneously concluding on an inferior treatment. The second noteworthy difference is that in the traditional setup an increase in replicates translates to an increase in the performance of the experimental findings. This is what leads to the reformulation of the problem where a practitioner typically determines the minimum number of replicates required to attain a pre-specified performance level. In our setting this requirement need not be specified. The cost associated with extended experimentation is implicitly built in as a cost of poorer performance which is equated with the gains. Finally, in the frequentist approach to DoE, a solution to on-line experimentation is proposed through EVolutionary OPeration (EVOP) by George E.P. Box (Box, 1957). The objectives of EVOP align with ours in that EVOP

seeks to facilitate *process improvement in the normal course of production*. This is synonymous with experimenting in an on-line environment. However, there are two significant differences between EVOP and our work. In terms of the environment, we assume statistical control, temporally, whereas EVOP seeks to adjust for an ever changing environment which could have drifts in the process. In terms of methodological strategies of minimizing the risk due to experimentation, EVOP adopts the approach of running traditional designed experiments over a narrower range. This, the authors note, could run the risk of failing to capture macroscopic trends, either owing to locally different behavior, or noise. Our approach, and those in traditional Bandit studies, explore the question of how much to experiment and on which treatments, for a fixed factor-level definition assuming a stable process.

An alternate line of inquiry is seen in the Bayesian studies that seek to determine the optimal sample size. Here, assumptions are made on the prior distribution of the treatments or the treatment differences. These techniques then adopt various criteria of performance, which incorporate both the costs and benefits of increased samples (Fraser and Guttman, 1956; Pham-Gia and Turkkan, 1992; Schlaifer and Raiffa, 1961; Schönbrodt et al., 2017; Tan and Machin, 2002; Willan and Pinto, 2005). The techniques then seek to identify sample sizes or dynamically make stopping rules for sampling based on this criterion. A study that typifies this approach is seen in Willan and Pinto (2005) which uses the idea of expected value of sample information (EVSI) Schlaifer and Raiffa (1961) to quantify the probabilistic value of experimenting against the costs associated with conducting experiments. Our work also broadly adopts the Bayesian approach and defines a criterion of cumulative improvement that we are maximizing. However, our work is different from Willan and Pinto (2005) and the other studies in that it is a specific implementation of the generic Bayesian approach for sample size determination to factorial designs. More importantly, since it is not a general framework for decision-making, the costs and benefits are explicitly defined (in the other studies that are referenced the costs and benefits are parameters which are inputs to the model). In other words, the online setting motivates us to not have any exogenous costs or benefits, but are instead built into the process of experimenting and exploiting. This, in turn, results in fewer parameters and enables us to create close-form solutions to identifying the optimal number of replicates.

Finally, the main and vast body of work related to sequential experimentation comes from the Bandit framework (Sutton and Barto, 1998). If the goal is pure exploration (offline environment) then this is captured in studies of best-arm identification (Audibert and Bubeck, 2010). A stronger parallel to the multi-factor DOE setting can be seen in the work of Soare (2015) which extends the sequential exploration through linear bandits and discusses it's connection to the G-optimal design. However, our problem statement seeks to balance exploration and exploitation in the online setting which is typified in the multi-armed bandit problem. Again the DOE multi-factor setting can be captured through the linear bandit conception and has been discussed in literature (Sudarsanam and Ravindran, 2017). However, an important distinction with these online learning approaches is the process of sequential decision-making, which restricts the experimental problem to non-parallel deployments only. Our work is built to exploit typical designed experiment settings which can, and sometimes must, be deployed in parallel.

3 Theory

In this section, we put forth the mathematical formulation of the problem statement and present theoretical results. We conceptualize a Bayesian framework to the General Linear Model with two-way and three-way interactions. We assume Gaussian priors for the co-efficients as presented in Joseph (2006). We then derive the expected value of improvement for various orthogonal experimental designs as a function of the number of replicates and the underlying environment. We then extend this cumulative improvement in the online context and identify the optimal number of replicates as a function of T, the total number of units which can be subject to the experimental settings.

The derivations for expected improvement are broadly along the lines of Frey and Wang (2006), with some notable extensions. The study of Frey and Wang (2006) was primarily concerned with resolution III arrays, which has the unique structure of two-way interactions being aliased with main effects, which in turn results in certain simplifications for the analysis of improvement. We extend this approach to all forms of factorial designs. Also, our work extends to three-way interactions, while the previous study was confined to two-way interactions. Finally, and most importantly, our major contribution lies in extending the derivation to cumulative improvement in the online context and deriving the optimal number of replicates which was not the focus of the previously mentioned study.

Similar to Frey and Wang (2006), our study seeks to determine the factor-levels by analyzing only the main effects, without any statistical significance tests. The authors of this study acknowledge that this can be a limitation, especially in environments where the interaction strength is high. While this analysis of experiments does not seek to explicitly model interaction terms, our study still captures the effect of interaction terms and quantifies the consequences of failing to exploit them. Therefore, if practitioners seek to use more advanced methods of analyzing data, the optimal r, which we derive in this section can be thought of as an lower bound. The underlying assumption here is that the other methods will lead to an improvement greater than or equal to those provided by an analysis that is limited to main effects. Also, in section 4 we explore computational results of explicitly modeling both main effects and interaction terms through regression, and compare these to the theoretical results developed in this section.

3.1 Mathematical Framework and Notations

$$y(x_1,x_2.....,x_k) = \sum_{i=1}^{k} \frac{\Delta_i}{2} x_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{\Delta_{ij}}{2} x_i x_j + \sum_{i=1}^{k-2} \sum_{j=i+1}^{k-1} \sum_{k=j+1}^{k} \frac{\Delta_{ijk}}{2} x_i x_j x_k + \varepsilon_k \qquad x_i \in \{1,-1\} \tag{1}$$

where

$y(x_1,x_2,...x_k)$ – Response for the given factors

x_i – Factor which takes value either 1 or -1

$\Delta_i \sim N(0,\sigma_m^2)$ – Coefficient of main effects
$\Delta_{ij} \sim N(0,\sigma_{2int}^2)$ – Coefficient of second order interaction effects
$\Delta_{ijk} \sim N(0,\sigma_{3int}^2)$ – Coefficient of third order interaction effects
$\varepsilon_k \sim N(0,\sigma_e^2)$ – Irreducible error

A list of related parameters

$Pr(L^*)$ – Probability of choosing the optimal level
EI_d – Expected Improvement for an experimental design d
CI_d – Cumulative Improvement for an experimental design d
r – Number of replicates
r^* – Optimal number of replicates
k – Number of factors
p – Degree of fractioning for a fractional factorial design
T – Total available number of units for exploration and exploitation
α – Likelihood of main effects being statistically significant
γ – Likelihood of second order interaction effects being statistically significant
ρ – Likelihood of third order interaction effects being statistically significant
$\phi = \frac{\sigma_m}{\sigma_e}$
$\psi = \frac{\sigma_{2int}}{\sigma_e}$
$\eta = \frac{\sigma_{3int}}{\sigma_e}$
N_s – Number of second order interaction effects which are confounded with a main effect
N_t – Number of third order interaction effects which are confounded with a main effect

3.2 Number of replicates for full factorial designs

Theorem 1 If a two level full factorial design with k factors is conducted in an online setup, where there are a total of T units and the environment is characterized by equation 1 with its corresponding priors, then the optimal number of replicates is given by:

$$r^* = \frac{-3+\sqrt{9+2T\phi^2}}{2^k\phi^2} \tag{2}$$

Proof In a 2^k design all three-way interactions and two-way interactions are completely orthogonal to the main effects. Hence any effort to estimate the main effect

co-efficients (or just their sign in order to make a decision) is unaffected by the presence of these terms. This leads to a simplification of equation 1. We rewrite this as:

$$y(x_1,x_2.....,x_k) = \sum_{i=1}^{k} \frac{\Delta_i}{2} x_i + \varepsilon_k \qquad x_i \in \{1,-1\} \tag{3}$$

where the assumptions and corresponding priors still hold.

Here, we choose a level $(-1 \text{ or } +1)$ for each factor independently. The probability of choosing the correct level is a function of whether our estimate of Δ_i ($\widehat{\Delta_i}$) is the same sign as Δ_i. In other words, the gap between the superior treatment to inferior treatment is represented through $|\Delta_i|$ which is a half-normal distribution. The probability of picking the superior treatment is if our estimate has the same sign, which can be captured by the probability that $N(|\Delta|, \frac{\sigma_e^2}{2^{k-2}r})$ takes on a value greater than 0. This is shown in the double integral below

$$Pr(L^*) = \int_0^\infty \int_0^\infty \frac{1}{\frac{\sigma_e}{\sqrt{2^{k-2}r}}\sqrt{2\pi}} exp(-\frac{(x-\Delta)^2}{\frac{2^{3-k}\sigma_e^2}{r}})\, dx \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp(-\frac{\Delta^2}{2\sigma_m^2})\, d\Delta \tag{4}$$

The outer integral (related to Δ) captures the probabilistic magnitude of the gap through the half-normal, where as the inner integral, over x, conveys the likelihood that the estimate of Δ is a positive value.

Then, the expected improvement after experimentation builds on the idea that with the probability $Pr(L^*)$ (equation 4), the experimenter will gain an improvement of $\frac{\Delta}{2}$, and with the $1-Pr(L^*)$ the experimenter will lose a value of $-\frac{\Delta}{2}$. The expected improvement (EI) after using $2 \times r$ units is formulated below:

$$\begin{aligned} EI_{2^k}(r) = \alpha \times k \times (&\int_0^\infty \int_0^\infty \frac{\Delta}{2} \frac{1}{\frac{\sigma_e}{\sqrt{2^{k-2}r}}\sqrt{2\pi}} exp(-\frac{(x-\Delta)^2}{\frac{2^{3-k}\sigma_e^2}{r}})\, dx \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp(-\frac{\Delta^2}{2\sigma_m^2})\, d\Delta \\ &+ \int_0^\infty \int_{-\infty}^0 -\frac{\Delta}{2} \frac{1}{\frac{\sigma_e}{\sqrt{2^{k-2}r}}\sqrt{2\pi}} exp(-\frac{(x-\Delta)^2}{\frac{2^{3-k}\sigma_e^2}{r}})\, dx \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp(-\frac{\Delta^2}{2\sigma_m^2})\, d\Delta) \end{aligned} \tag{5}$$

$$= \frac{\alpha k \sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{2^{2-k}}{r}\frac{\sigma_e^2}{\sigma_m^2}}} \tag{6}$$

if we write $\phi = \frac{\sigma_m}{\sigma_e}$ then

$$EI_{2^k}(r) = \frac{\alpha k \sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{2^{2-k}}{r}\frac{1}{\phi^2}}} \tag{7}$$

The equation above captures the expected improvement post experimentation which can be exploited on $(T-2^k r)$ units and the expected improvement during the experimental phase (of $2^k r$ units) is zero. The cumulative improvement after T units is given by:

$$CI_T = (T-2^k r)\frac{\alpha k \sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{2^{2-k}}{r}\frac{1}{\phi^2}}} \tag{8}$$

To find the maxima, we differentiate equation 8 with respect to r and equate it to zero.

$$\frac{\partial((T-2^k r)\frac{\alpha k \sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{2^{2-k}}{r}\frac{1}{\phi^2}}})}{\partial r} = 0 \tag{9}$$

The main claim of Theorem 1 as shown in Equation 2 follows from simplifying the equation above.

3.3 Number of replicates for fractional factorial designs

Table 1: Optimal number of replicates for fractional factorial designs

Design Specification	Expected Improvement	Theoretical r^*
2^{3-1}_{III}	$\frac{3\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\gamma\frac{\psi^2}{\phi^2}+\frac{1}{r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T(\phi^2+\gamma\psi^2)}}{4(\phi^2+\gamma\psi^2)}$
2^{4-1}_{IV}	$\frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\rho\frac{\eta^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T(\phi^2+\rho\eta^2)}}{8(\phi^2+\rho\eta^2)}$
2^{5-1}_{V}	$\frac{5\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{4r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T\phi^2}}{16\phi^2}$
2^{5-2}_{V}	$\frac{\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+2\gamma\frac{\psi^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}}+\frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\gamma\frac{\psi^2}{\phi^2}+\rho\frac{\eta^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}}$	** [2]
2^{6-1}_{VI}	$\frac{6\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{8r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T\phi^2}}{32\phi^2}$
2^{6-2}_{IV}	$\frac{6\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+2\rho\frac{\eta^2}{\phi^2}+\frac{1}{4r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T(\phi^2+2\rho\eta^2)}}{16(\phi^2+2\rho\eta^2)}$
2^{6-3}_{III}	$\frac{6\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+2\gamma\frac{\psi^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T(\phi^2+2\gamma\psi^2)}}{8(\phi^2+2\gamma\psi^2)}$
2^{7-1}_{VII}	$\frac{7\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{16r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T\phi^2}}{64\phi^2}$
2^{7-2}_{IV}	$\frac{3\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{8r}\frac{1}{\phi^2}}}+\frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\rho\frac{\eta^2}{\phi^2}+\frac{1}{8r}\frac{1}{\phi^2}}}$	**2
2^{7-3}_{IV}	$\frac{7\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+4\rho\frac{\eta^2}{\phi^2}+\frac{1}{4r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T(\phi^2+4\rho\eta^2)}}{16(\phi^2+4\rho\eta^2)}$
2^{7-4}_{III}	$\frac{7\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+3\gamma\frac{\psi^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}}$	$\frac{-3+\sqrt{9+2T(\phi^2+3\gamma\psi^2)}}{8(\phi^2+3\gamma\psi^2)}$

The proof for theorems 2 through 6 are presented in the Appendix.

[2] Owing to space constraints these formulations are excluded. They can be found in the Appendix

Unlike the full factorial designs, the optimal number of replicates for fractional factorial arrays do not have a general form. This is owing to the fact that different designs have different resolutions and aliasing patterns. We study all possible fractional factorial arrays for designs with 7 factors or less.

Theorems 2 - 6 present the expected improvement for each design after r replicates and also derive an optimal number of replicates for each design. We present these in Table 1 below

3.4 Inferences from the theoretical proofs

There are three broad inferences we can draw from the proofs developed in sections 3.2 and 3.3:

1. An important driver of improvement and the optimal level of experimentation is based on ϕ or the ratio $\frac{\sigma_m}{\sigma_e}$. A larger σ_m (for a given σ_e) translates to greater absolute gains, since the gap between a random treatment (or the average treatment) and the optimal treatment is higher. More interestingly, we can observe that across the board a higher ϕ translates to a lower requirement of experimentation (in terms of replicates) for a given design. This is primarily driven by two complementary reasons. First, when ϕ is high, a given level of statistical certainty on the optimal treatment is arrived at with fewer experiments. Second, the cost of experimentation is higher, since the balanced experiments are subjecting resources to a suboptimal treatment which differs from the optimal one by a larger magnitude (since σ_m is relatively high). The parameter α related to likelihood of significance has no impact on the optimal level of experimenting since it only scales up (or down) the cumulative improvement by a fixed factor.
2. An increase in the strength and likelihood of interaction effects, captured by ψ, η, ρ, and γ leads to a decrease in the overall improvement. This is because a classical analysis is not capable of exploiting these effects, and their role could probabilistically lead to inaccurate inferences on main effects, in designs where there is confounding of the main effect with the interaction terms. An increase in these effects also lead to a decrease in the optimal number of replicates for the designed experiment. This is owing to the fact that for a given system and design, when these parameters are high, the dividends from conducting experiments are more gradual with increasing replicates. In other words, one gains less from increasing the replicates.
3. Finally, we observe that in designs where there is no confounding of main effects with two and three-level interactions, the optimal level of experimentation is the same. This might seem contrary to the results, since smaller designs seem to require more replicates. However, this scales inversely with the number of treatment combinations, leading to the exact same number of experimental units to be expended. A closer look at equation 2 for full factorials would serve as a validation. This can also be extended to compare different levels of fractions for designs with the same number of factors. For example, we could consider the comparison between the 2^7 and the 2^{7-1}_{VII} (both designs do not confound main effects with two-way and three-way interactions). However, when this is extended to fractions

with confounding (for instance 2_{IV}^{7-2}, 2_{IV}^{7-3} and 2_{III}^{7-4}), we observe that due to the effect discussed in inference 2, the designer would be motivated to experiment less.

4 Empirical Validation and numerical results

In this section, we validate our theoretical findings with simulations conducted on a meta model of real world experiments and present indicative numerical results for various environments. We use the data set of 113 published experiments studied in the Li et al. (2006). We adopt the parameters and regularities as identified by the same study. We then build a probabilistic meta model as described in Frey and Li (2008). We use this to simulate response surfaces, as well as the noise in the experimental environment. We then study on-line performance, empirically, from using a different number of replicates. Each iteration fixes a given response surface, and then generates the maximum number of replicates with noise. This regulates the uncertainty due to responses, and replicates, to the minimum required while simultaneously exploring the inherent variability in using different response surfaces across iterations. In essence, the difference in average performance is accurately captured while minimizing variability due to the generation of pseudo random numbers.

Our empirical study seeks to (i) Validate the theoretical findings through a simulated environment that reflects the regularities of sparsity, hierarchy and also heredity (the last regularity is not captured in the theoretical model), (ii) Explore the effect of explicitly modeling interactions through a regression based approach, (iii) Establish baseline improvements that can be attained from each method of analysis and use it to quantify the value that can be gained through such a study, and finally,(iv) Provide indicative numerical results of the optimal number of replicates across various environments which can be easily used by practitioners. The topics (i), (ii) and (iii) are studied in section 4.1, and topic (iv) is studied in section 4.2.

4.1 The effect of environmental regularities and using regression based approaches on improvement

The empirical explorations in this section are three-fold. The first is to use the hierarchical probability model (HPM) as a simulation based empirical study to validate our theoretical results. The parameters derived from real-world data and the characterization of regularities such as *sparsity*, *hierarchy*, and *heredity* are adopted from Frey and Li (2008). Second, we consider the effect of explicitly modeling interaction terms through a regression based approach, which is a more common, and statistically sound approach to modeling and optimization. We seek to understand the major differences in performance through such an approach. Third, we establish a baseline of performance that can be expected through experimentation that is uninformed by our studies. This allows us to compare the improvement witnessed when the optimal number of replicates are chosen with the baseline to quantify the value of using such an approach. In this section, as an illustrative exercise we focus on designs with 7

factors. In particular, we study the five designs of the 2^7 full factorial, and the four fractional factorials 2^{7-1}_{VII}, 2^{7-2}_{IV}, 2^{7-3}_{IV} and 2^{7-4}_{III}. We confine our analysis to the case where the total number of units (T) is 10,000 and to a value of $\phi = 1$. In the numerical results summarized in section 4.2 we explore various values of ϕ and T.

We assert that the empirical study of 2^7 full factorial or 2^{7-1}_{VII} fractional factorial design, when used in conjunction with classical analysis, serves as true validation of the theoretical results. Whereas, the empirical analysis of the remaining three fractional factorial designs can broadly serve to quantify any differences in performance due to an environmental regularity termed as *heredity* which is captured in the simulations. The environmental regularity, *heredity*, captures the phenomena that interaction effects are conditionally more likely to be significant if their parent effects are significant. While heredity would play no role in the 2^7 full factorial or 2^{7-1}_{VII} fractional factorial design (since there is no confounding of main effects with two-way or three-way interactions), in the other designs it could potentially cause a disparity in actual performance from the theoretical solution. This is owing to the fact that our theoretical analysis determines γ and ρ as expected values of the likelihood of significance and the more intricate dependence structure could cause difference in actual performance. However, this is expected to marginal, given the *sparsity* of two and three way interactions, as well as the strength owing to *hierarchy*. Figures 1, 2,3,4 and 5 captures these studies across the five 7 factor designs.

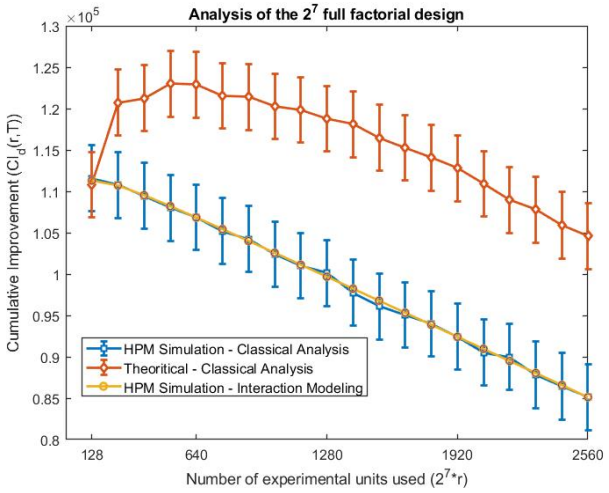

Fig. 1: Cumulative improvement for 2^7 design for varying numbers of experimental samples out of T = 10,000 units and ϕ=1

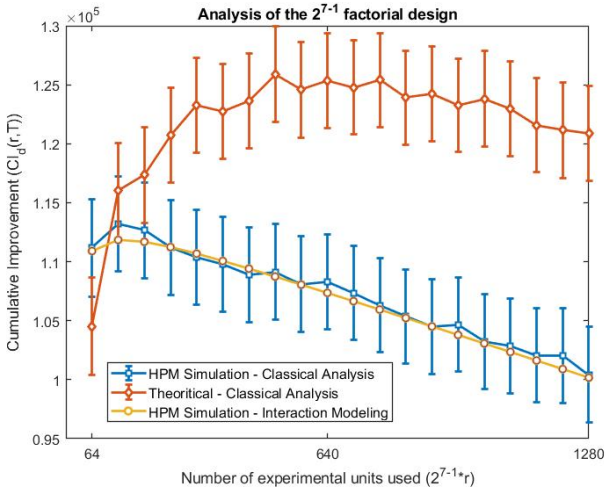

Fig. 2: Cumulative improvement for 2^{7-1}_{VII} design for varying numbers of experimental samples out of T = 10,000 units and ϕ=1.0

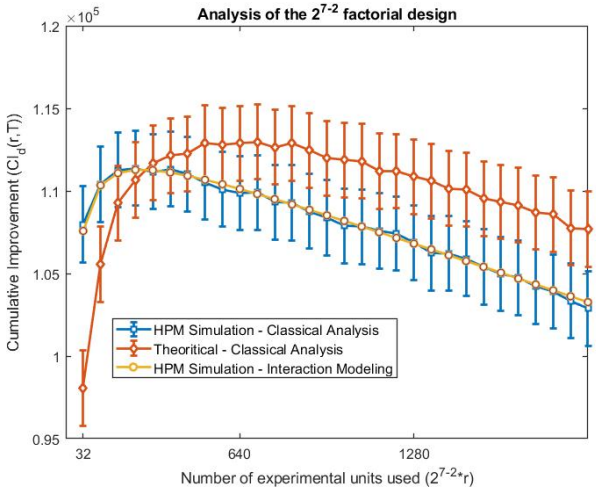

Fig. 3: Cumulative improvement for 2^{7-2}_{IV} design for varying numbers of experimental samples out of T = 10,000 units and ϕ=1.0

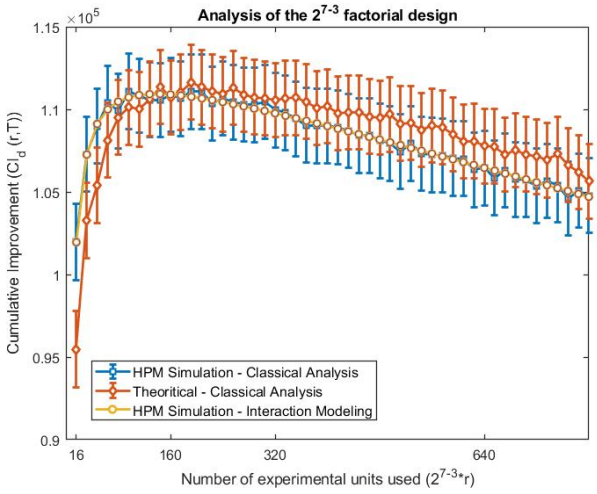

Fig. 4: Cumulative improvement for 2^{7-3}_{IV} design for varying numbers of experimental samples out of T = 10,000 units and ϕ=1

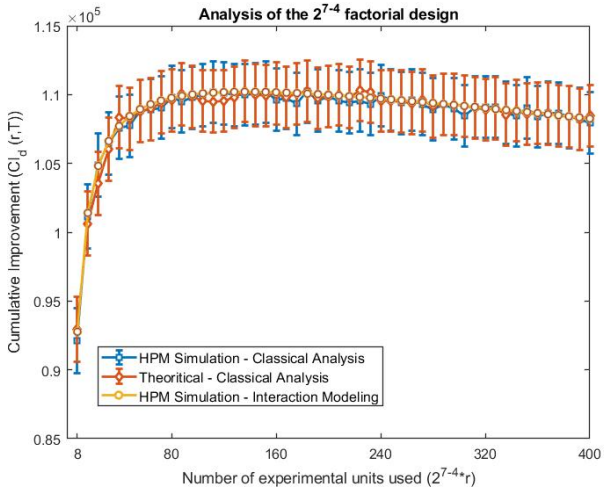

Fig. 5: Cumulative improvement for 2^{7-4}_{III} design for varying numbers of experimental samples out of T = 10,000 units and ϕ=1

At a high-level, the results indicate that the cumulative improvement from the HPM based simulation with classical analysis is statistically in-line with the theoretical results for the 2^7 full factorial or 2^{7-1}_{VII} fractional factorial design. In the cases of

the 2_{IV}^{7-2}, 2_{IV}^{7-3} and 2_{III}^{7-4} designs, where *heredity* could cause a difference between the theoretical results and the empirical performance, we observe that this effect is too small for the environments studied. The largest difference is seen in the 2_{IV}^{7-2} and to significantly lower degrees in 2_{IV}^{7-3} and 2_{III}^{7-4}. Also, any minor differences in performance between the theoretical and HPM based classical analysis in these designs show no difference to cause an impact on the optimal number of replicates recommended by both approaches (they both peak at the same number of replicates).

With respect to the use of a regression-based interaction modeling, we find that the design matrix plays a significant role. The larger design matrices (full fractions and high resolution designs), are capable of exploiting more parameters in the model, and this has two effects. First, the overall improvement using regression-based interaction modeling is superior to classical analysis, for the optimal number of replicates. The 2^7 full factorial or 2_{VII}^{7-1} fractional factorial have the biggest difference in performance (they can both exploit all two-way and three-way interactions), where as the 2_{III}^{7-4} results in an identical performance to classical analysis since the interaction modeling approach cannot model any interaction terms. Second, with the larger designs where there is a significant improvement in regression based modeling, the optimal number of replicates is higher. In the 2_{VII}^{7-1}, 2_{IV}^{7-2}, and 2_{IV}^{7-3} designs, a low number of replicates results in the interaction modeling approach having poorer performance than classical analysis. This is primarily because of overfitting the model to limited data. By contrast, the classical analysis is a high bias, low variance model that is robust to the noise in the system. However, with more data (more replicates), the interaction modeling approach is capable of utilizing the information to create a more refined model. Whereas, we can see that classical analysis with its limitations of exploiting only the main effects is limited in improvement. As a summary, with larger design matrices, we can conclude that the optimal number of replicates is higher than the theoretical results if one chooses to analyze the data through a regression based approach which explicitly models the interaction terms. The theoretical results can therefore be used as a lower bound requirement on the experimentation needed in these cases.

In order to provide a relative, numeric measure of the optimal cumulative improvement seen in figures 1 through 5, we establish a baseline performance. Here, we assume that a practitioner, uninitiated to such a study, could choose to experiment a uniformly random percentage of the total available units. We establish the baseline as the expected value of the cumulative improvement that can be gained when the number of replicates is a uniform random variable. This is compared to the cumulative improvement of the proposed method which is the value obtained when the optimal number of replicates are selected. Having established that theoretical results with the classical analysis are in line with the empirical studies, we present the percentage improvement over the baseline only for the theoretical classical analysis and the empirical modeling with regression based interaction modeling.

Table 2: Percentage improvement in cumulative improvement from deploying the optimum number of replicates recommended in the proposed framework over the Baseline of a random selection of replicates

Design	Classical Analysis (Theoretical)			Regression Based Interaction modeling		
	Baseline (Expected Cumulative Improvement with a random number of replicates) (no units)	Proposed Method (Cumulative Improvement with optimal number of replicates) (no units)	Percentage Improvement over Baseline [1]	Baseline (Expected Cumulative Improvement with a random number of replicates) (no units)	Proposed Method (Cumulative Improvement with optimal number of replicates) (no units)	Percentage Improvement over Baseline [3]
2^7	57,083	111,315	95.01%	67,725	122,989	81.61%
2^{7-1}_{VII}	57,083	111,315	95.01%	67,872	125,903	85.51%
2^{7-2}_{IV}	57,068	111,287	95.01%	58,895	112,962	91.81%
2^{7-3}_{IV}	57,057	111,266	95.01%	57,218	111,633	95.12%
2^{7-4}_{III}	56,569	110,337	95.05%	56,483	110,313	95.30%

The results indicate that the cumulative improvement that can be expected through the recommended number of replicates are at least 81.6% more than a random guess of degree of experimentation, and can be at most 95.3%, across the designs explored in this section and the two different methods of analysis. This magnitude of improvement cannot directly be inferred from figures 1 through 5, because the figures explore a fairly narrow range of potential replicates close to the optima, whereas a truly uninformed random guess leads to a substantially lower expected value of cumulative improvement.

4.2 Numerical results for optimal sample size under various environments

In this section, we present numerical results for various environments. This can be conveniently adopted by a practitioner who is looking for indicative values. Our study explores values of $\phi = 0.5, 1, 2$ and population sets of $T = 100$, 1000 and 10000. Similar to table 2, we only compare the theoretical classical analysis and the regression based interaction modeling. This is owing to the fact that the theoretical results with the classical analysis are in line with the empirical studies. We set $\alpha = 0.41$, which is adopted from the findings of Li et al. (2006) and accounts for *sparsity* of effects (Box and Hunter, 1986), a documented pattern of the experimental environment which states that a subset of the total input parameters in an experiment tend to be significant (have a real effect on the output). We fix η and ψ, based on the relative magnitude of σ_{2int} and σ_{3int} as determined by Frey and Li (2008). This is done in order to capture a regularity termed as *hierarchy* (Hamada and Wu, 1992), which captures the idea that main effects tend to have a higher magnitude than two-way interactions, which in turn tend to have a higher magnitude than three-way interactions and so on. For instance, a $\phi = \frac{\sigma_m}{\sigma_e} = 1.0$ would result in a $\psi = \frac{\sigma_{2int}}{\sigma_e} = 0.278$ and $\eta = \frac{\sigma_{3int}}{\sigma_e} = 0.137$. Also, similar to the use of α to capture the likelihood of a main effect being statistically significant, we use the terms γ and ρ to signify the likelihood of two-way and three-way interactions being significant, respectively. This extends the idea of *sparsity* to interaction terms. The values of $\gamma = 0.079$ and $\rho = 0.048$ are also adopted

[3] The improvement is defined as the percentage gain of the proposed method over the baseline in absolute terms, since the lowest possible improvement is 0, which occurs at 0 replicates and maximum possible replicates

from the same source (Li et al., 2006). In Table 3, we consider all factorial designs with 7 factors or less.

Table 3: Indicative numerical results: A practitioners cheat sheet for optimal number of replicates

Design Specification	No of units	Optimal r^* through classical analysis			Optimal r^* through regression		
		$\phi = 0.5$	$\phi = 1$	$\phi = 2$	$\phi = 0.5$	$\phi = 1$	$\phi = 2$
2^2	100	5	3	2	6	4	2
	1000	20	11	6	24	13	7
	10000	68	35	18	78	40	21
2^3	100	3	2	1	4	3	2
	1000	10	6	3	13	8	4
	10000	34	18	9	42	22	11
2^{3-1}_{III}	100	5	3	2	5	3	2
	1000	19	11	6	20	11	6
	10000	67	35	18	66	35	17
2^4	100	2	1	1	3	2	1
	1000	5	3	2	7	4	3
	10000	17	9	5	24	13	7
2^{4-1}_{IV}	100	3	2	1	4	2	1
	1000	10	6	3	12	7	4
	10000	34	18	9	41	22	11
2^5	100	1	1	1	2	2	1
	1000	3	2	1	5	3	2
	10000	9	5	3	15	8	5
2^{5-1}_{V}	100	2	1	1	3	1	1
	1000	5	3	2	7	4	3
	10000	17	9	5	22	12	7
2^{5-2}_{III}	100	3	2	1	3	2	1
	1000	10	6	3	11	6	3
	10000	34	18	9	36	19	9
2^6	100	1	1	1	2	2	1
	1000	2	1	1	3	2	1
	10000	5	3	2	8	6	3
2^{6-1}_{VI}	100	1	1	1	2	2	2
	1000	3	2	1	6	3	2
	10000	9	5	3	17	11	6
2^{6-2}_{IV}	100	2	1	1	2	2	1
	1000	5	3	2	6	4	2
	10000	17	9	5	20	11	6
2^{6-3}_{III}	100	3	2	1	3	2	1
	1000	10	6	3	10	6	3
	10000	34	18	9	34	18	9
2^7	100	N/A	N/A	N/A	N/A	N/A	N/A
	1000	1	1	1	2	1	1
	10000	3	2	1	6	3	2
2^{7-1}_{VII}	100	1	1	1	2	2	1
	1000	2	1	1	4	3	2
	10000	5	3	2	11	7	4
2^{7-2}_{IV}	100	1	1	1	2	2	1
	1000	3	2	1	4	3	2
	10000	9	5	3	16	10	6
2^{7-3}_{IV}	100	2	1	1	3	2	2
	1000	5	3	2	7	5	4
	10000	17	9	5	23	14	9
2^{7-4}_{III}	100	3	2	1	3	2	1
	1000	10	6	3	10	6	3
	10000	33	18	9	33	18	9

5 Conclusions and future work

The study looks at determining the optimal expenditure for one-shot, non-sequential experimental plans, which are typified by A/B tests, split tests, pilot studies, randomized control trials, statistical designed experiments, etc., in an online environment. We illustrate this work through theoretical derivations of the optimal number of replicates for full and fractional factorial arrays with two to seven factors. The novelty in our approach stems from our assessment through cumulative improvement, which is suited for the online environment, rather than traditional offline statistical metrics (OC, curves, p-values, lift scores). Our empirical explorations through simulations validate the theoretical results and also facilitate the exploration of specific environments and techniques of analysis that are not supported by the theory. Our conclusions from this study are broadly divided across insights pertaining to the environment and those relating to the method of analysis. Both of these play an important role in determining the degree to which a practitioner should commit resources to experiments.

With respect to the environment, we observer that the most important parameters are the total number of units (T) and ϕ (the ratio $\frac{\sigma_m}{\sigma_e}$). With respect to T, we find that a larger value requires more experimentation in absolute terms, because there are more units at stake once a decision is made. However, if we study percentage of experimentation (which for full factorials would be defined as $\frac{r^* \times 2^k}{T}$), then a larger T requires a smaller percentage for experimentation since the increase in statistical certainty asymptotically reduces beyond a certain number of experiments. With respect to the ranges studied, for the value of $\phi = 1$, the percentage of experimentation ($\frac{r^* \times 2^k}{T}$) ranges from 11% for 100 units to 1% for 10,000 units. With respect to ϕ, a larger value results in lower experimentation. As an indication, for full factorial designs, when T is set to 1000 units, $\phi = 0.5$ results in 7.82% of experimentation whereas $\phi = 2$ results in 2.16%. We see the most requirement for experimentation (as a percentage) when T and ϕ are both low. at $T = 100$ and $\phi = 0.5$ we require 20%. Conversely, when $T = 10{,}000$ and $\phi = 2$ we need less than 1%. Given the assumptions of sparsity and hierarchy adopted from earlier studies, the roles of ψ, η, ρ, and γ are minimal in the theoretical analysis. However, the ability of the regression based approaches to leverage interaction effects suggests that degree of experimentation could be more sensitive to these terms using this approach.

With respect to the method of analysis, we observe that regression based interaction modeling reaches optimal performance with more replicates than a classical analysis for full factorial and high resolution designs. This is because regression based approaches that explicitly model interactions have more to gain from extended experimentation as opposed to classical analysis when the interaction terms can be statistically inferred. Adopting values for *strength* and *likelihood* of interactions relative to the main effects from Li et al. (2006), we find that the requirement for experimentation increases by approximately 90% to 120% when interaction modeling through regression is used as opposed classical analysis. With low-resolution designs which can only model a subset of the interactions or cannot model interactions fully (due to there being insufficient degrees of freedom), the difference between the two ap-

proaches decreases commensurately. In designs where no interaction terms can be modeled (like the 2^{7-4}_{III}) both approaches yield identical recommendations for cumulative improvement as well as the optimal level of experimentation. Across both types of analyses, we observe that performance at the optimal number of replicates is significantly greater than the expected improvement for a random guess of replicates. This difference is consistently equal to a 95% for the classical analysis, and can vary between 82% to 95% for the interaction modeling.

In terms of future work, there exists a wide range of possible extensions that one could explore. One such extension, which could substantially increase the real-world applicability, is to consider quasi-sequential environments. In this study, we consider a one-shot experimental design followed by implementation. This is essentially a two-phase setup– one for exploration and the other for exploitation. By contrast, the bandit studies follow a fully sequential form of one-unit-at-a-time (there are as many phases as units). There could be many real-world settings which can afford some number of phases in between. For instance, if the time horizon and cost structure could support three separate phases (or batches), what should be the resource allocation towards each phase? More interestingly in the middle phase what should be the ratio of allotment of treatments to resources (which understandably need not be balanced, as it is for the first phase, or an entire commitment to the preferred treatment, as it is in the last phase).

Another area of inquiry could be to look at alternate distributions that represent the true mean of treatments, as well as the noise distribution. Future studies could also explore environments which are distribution-free or have limited assumptions associated with distributions. In alignment with the major focus of DoE literature, all the designs explored in this study consider 2-level designs. Another prospective extension could be the exploration of experiments with 3 or more level designs, which we believe goes beyond a mathematically trivial extension of the 2-level case. The second prospective area for future work could surround analyses that are not restricted to maximizing the expected cumulative improvement. This work could model the entire distribution of the outcome variable conditioned on the controllable ones. One could also look at modeling of upper/lower bounds in performance, which naturally aligns with using fewer distribution-related assumptions as priors. We see this in the use Chernoff bounds which are adopted extensively in the Bandit literature. These extensions would also help align the decision-making process with the risk profile of the practitioner or context.

Acknowledgements This work was partially supported by the Robert Bosch Center for Data Science and Artificial Intelligence (RBC-DSAI) at IIT Madras.

Appendix 1: Proofs

Theorem 2:[4]

This theorem shows the derivations for all arrays that have a resolution V or higher. Here, the main effects are only confounded with four-way interactions or higher. The optimal number of replicates is given by:

$$r^* = \frac{-3+\sqrt{9+2T\phi^2}}{2^{k-p}\phi^2} \tag{A.1}$$

Proof:
All fractional factorial arrays with a resolution V or higher can be captured in this derivation. Some examples are: 2_V^{5-1}, 2_{VI}^{6-1} and 2_{VII}^{7-1} . In all these designs the main effects are aliased with four way interactions or higher.The estimate of Δ for these designs follows $N\big(|\Delta|,\big(\frac{\sigma_e^2}{2^{k-p-2}r}\big)\big)$ and the expected improvement at an optimal level is:

$$EI(r) = k\times\alpha\times\Big(\int_0^\infty\int_0^\infty \frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\Big(\frac{\sigma_e^2}{2^{k-p-2}r}\Big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(\frac{\sigma_e^2}{2^{k-p-2}r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma_m^2})\,d\Delta$$
$$+\int_0^\infty\int_{-\infty}^0 -\frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\Big(\frac{\sigma_e^2}{2^{k-p-2}r}\Big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(\frac{\sigma_e^2}{2^{k-p-2}r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma_m^2})\,d\Delta\Big) \tag{A.2}$$

$$= \frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{2^{k-p-2}r}\frac{\sigma_e^2}{\sigma_m^2}}} \tag{A.3}$$

Let $\frac{\sigma_m}{\sigma_e}=\phi$, $\frac{\gamma\sigma_{2int}}{\sigma_e}=\psi$ and $\frac{\rho\sigma_{3int}}{\sigma_e}=\eta$

$$= \frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{2^{k-p-2}r}\frac{1}{\phi^2}}} \tag{A.4}$$

The equation A.4 gives the expected improvement for the (T-2^{k-p}r) units and the expected improvement during the experiment (nr) is zero. The cumulative improvement after T units is given by:

$$CI_T = (T-2^{k-p}r)\frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{2^{k-p-2}r}\frac{1}{\phi^2}}} \tag{A.5}$$

To find the maxima, differentiate the above function with respect to r and make it equal to zero.

$$\frac{\partial CI_T}{\partial r} = 0 \tag{A.6}$$

Solve the resultant equation to get the closed form equation for optimal number of replicates

$$r^* = \frac{-3+\sqrt{9+2T\phi^2}}{2^{k-p}\phi^2} \tag{A.7}$$

Theorem 3

This theorem covers standard resolution III designs (the one exception is the 2^{5-2} which is covered in Theorem 5). Here, the main effects are confounded with the two-way interaction effects. The optimal number of replicates is given by:

[4] The proof for theorem 1 is in the main body of the manuscript - section 3

$$r^* = \frac{-3+\sqrt{9+2T(\phi^2+N_s\gamma\psi^2)}}{2^{k-p}(\phi^2+N_s\gamma\psi^2)} \tag{A.8}$$

Proof:
Examples of designs that are covered by this theorem are: 2^{3-1}_{III},2^{6-3}_{III} and 2^{7-4}_{III}. Here, each main effect in the 2^{3-1}_{III} design is confounded with one two-way interaction effect. Whereas, each main effect in 2^{6-3}_{III} design is confounded with 2 two-way interaction effects. Finally, each main effect in 2^{7-4}_{III} design is confounded with 3 two-way interaction effects. The estimate of Δ for these designs follows $N\big(|\Delta|,\big(N_s\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)\big)$ and the expected improvement at an optimal level is given by

$$EI(r) = k\times\alpha\times\Big(\int_0^\infty\int_0^\infty \frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\big(N_s\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(N_s\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma^2_m})d\Delta$$
$$+\int_0^\infty\int_{-\infty}^0 -\frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\big(N_s\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(N_s\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma^2_m})d\Delta\Big) \tag{A.9}$$

$$= \frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+N_s\gamma\frac{\sigma^2_{2int}}{\sigma^2_m}+\frac{1}{2^{k-p-2}r}\frac{\sigma^2_e}{\sigma^2_m}}} \tag{A.10}$$

Let $\frac{\sigma_m}{\sigma_e}=\phi$, $\frac{\sigma_{2int}}{\sigma_e}=\psi$ and $\frac{\sigma_{3int}}{\sigma_e}=\eta$

$$= \frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+N_s\gamma\frac{\psi^2}{\phi^2}+\frac{1}{2^{k-p}r}\frac{1}{\phi^2}}} \tag{A.11}$$

The cumulative improvement after T units is given by:

$$CI_T = (T-2^{k-p}r)\frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+N_s\gamma\frac{\psi^2}{\phi^2}+\frac{1}{2^{k-p}r}\frac{1}{\phi^2}}} \tag{A.12}$$

To find the maxima, differentiate the above function with respect to r and make it equal to zero.

$$\frac{\partial CI_T}{\partial r} = 0 \tag{A.13}$$

$$r^* = \frac{-3+\sqrt{9+2T(\phi^2+N_s\gamma\psi^2)}}{2^{k-p}(\phi^2+N_s\gamma\psi^2)} \tag{A.14}$$

Theorem 4

This theorem covers standard resolution *IV* designs (the 2^{7-2} is an exception that is covered in theorem 6). Here, the main effects are confounded with the third order interaction effects. The optimal number of replicates is given by:

$$r^* = \frac{-3+\sqrt{9+2T(\phi^2+N_t\rho\eta^2)}}{2^{k-p}(\phi^2+N_t\rho\eta^2)} \tag{A.15}$$

Proof:
Examples of designs that are covered by this theorem are: 2^{4-1}_{IV}, 2^{6-2}_{IV} and 2^{7-3}_{IV}. Here, each main effect in the 2^{4-1}_{IV} design is confounded with one three-way interaction effect. Whereas, each main effect in 2^{6-2}_{IV} design is confounded with 3 three-way interaction effects. Finally, each main effect in 2^{7-3}_{IV}

design is confounded with 4 three-way interaction effects. The estimate of Δ for these designs follows $N\big(|\Delta|,\big(N_t\rho\sigma^2_{3int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)\big)$ and the expected improvement at an optimal level is given by:

$$EI(r)=k\times\alpha\times\Big(\int_0^\infty\int_0^\infty\frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\big(N_t\rho\sigma^2_{3int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(N_t\rho\sigma^2_{3int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma^2_m})\,d\Delta$$
$$+\int_0^\infty\int_{-\infty}^0-\frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\big(N_t\rho\sigma^2_{3int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(N_t\rho\sigma^2_{3int}+\frac{\sigma^2_e}{2^{k-p-2}r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma^2_m})\,d\Delta\Big) \tag{A.16}$$

$$=\frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+N_t\rho\frac{\sigma^2_{3int}}{\sigma^2_m}+\frac{1}{2^{k-p-2}r}\frac{\sigma^2_e}{\sigma^2_m}}} \tag{A.17}$$

Let $\frac{\sigma_m}{\sigma_e}=\phi$, $\frac{\sigma_{2int}}{\sigma_e}=\psi$ and $\frac{\sigma_{3int}}{\sigma_e}=\eta$

$$=\frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+N_t\rho\frac{\eta^2}{\phi^2}+\frac{1}{2^{k-p-r}r}\frac{1}{\phi^2}}} \tag{A.18}$$

The cumulative improvement after T units is given by:

$$CI_T=(T-2^{k-p}r)\frac{k\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+N_t\rho\frac{\eta^2}{\phi^2}+\frac{1}{2^{k-p-r}r}\frac{1}{\phi^2}}} \tag{A.19}$$

To find the maxima, differentiate the above function with respect to r and make it equal to zero.

$$\frac{\partial CI_T}{\partial r}=0 \tag{A.20}$$

$$r^*=\frac{-3+\sqrt{9+2T(\phi^2+N_s\gamma\psi^2)}}{2^{k-p}(\phi^2+N_s\gamma\psi^2)} \tag{A.21}$$

Theorem 5: Optimal number of replicates for 2^{5-2}_{III} design

The confounding structure is not same for all main effects like previous designs. Among five main effects, one main effect is confounded with two way and five way interaction effects and the remaining main effects are confounded with two way, three way and four way interactions effects. The expected improvement derivation is not same for all main effects since the confounding structure is not same for all main effects.

Therefore, the estimate of Δ which is confounded with two way and five way interaction effects follows $N\big(|\Delta|,\big(2\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2r}\big)\big)$ and the expected improvement at an optimal level is given by

$$EI_{2^{5-2}_V}(r)=\alpha\times\int_0^\infty\int_0^\infty\frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\Big(2\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2r}\Big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(2\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma^2_m})\,d\Delta$$
$$+\int_0^\infty\int_{-\infty}^0-\frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\Big(2\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2r}\Big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(2\gamma\sigma^2_{2int}+\frac{\sigma^2_e}{2r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma^2_m})\,d\Delta \tag{A.22}$$

$$= \frac{\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+2\gamma\frac{\sigma_{2int}^2}{\sigma_m^2}+\frac{1}{2r}\frac{\sigma_e^2}{\sigma_m^2}}} \tag{A.23}$$

Similarly, the estimate of other Δs for 2^{5-1} design follows $N\big(|\Delta|,\big(\rho\sigma_{3int}^2+\gamma\sigma_{2int}^2+\frac{\sigma_e^2}{2r}\big)\big)$ and the expected improvement at an optimal level is

$$EI_{2_V^{5-2}}(r) = 4\times\alpha\times\Bigg(\int_0^\infty\int_0^\infty \frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\Big(\rho\sigma_{3int}^2+\gamma\sigma_{2int}^2+\frac{\sigma_e^2}{2r}\Big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(\rho\sigma_{3int}^2+\gamma\sigma_{2int}^2+\frac{\sigma_e^2}{2r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma_m^2})d\Delta$$
$$+\int_0^\infty\int_{-\infty}^0 -\frac{\Delta}{2}\ \frac{1}{\sqrt{2\pi\Big(\rho\sigma_{3int}^2+\gamma\sigma_{2int}^2+\frac{\sigma_e^2}{2r}\Big)}}exp\Big(-\frac{(x-\Delta)^2}{2\big(\rho\sigma_{3int}^2+\gamma\sigma_{2int}^2+\frac{\sigma_e^2}{2r}\big)}\Big)dx\frac{\sqrt{2}}{\sigma_m\sqrt{\pi}}exp(-\frac{\Delta^2}{2\sigma_m^2})d\Delta\Bigg) \tag{A.24}$$

$$= \frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\gamma\frac{\sigma_{2int}^2}{\sigma_m^2}+\rho\frac{\sigma_{3int}^2}{\sigma_m^2}+\frac{1}{2r}\frac{\sigma_e^2}{\sigma_m^2}}} \tag{A.25}$$

The expected improvement for 2^{5-2} design is

$$= \frac{\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+2\gamma\frac{\sigma_{2int}^2}{\sigma_m^2}+\frac{1}{2r}\frac{\sigma_e^2}{\sigma_m^2}}} + \frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\gamma\frac{\sigma_{2int}^2}{\sigma_m^2}+\rho\frac{\sigma_{3int}^2}{\sigma_m^2}+\frac{1}{2r}\frac{\sigma_e^2}{\sigma_m^2}}} \tag{A.26}$$

Let $\frac{\sigma_m}{\sigma_e}=\phi$, $\frac{\sigma_{2int}}{\sigma_e}=\psi$ and $\frac{\sigma_{3int}}{\sigma_e}=\eta$

$$= \frac{\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+2\gamma\frac{\psi^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}} + \frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\gamma\frac{\psi^2}{\phi^2}+\rho\frac{\eta^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}} \tag{A.27}$$

The cumulative improvement after T units is given by:

$$\underset{r}{\text{Max}}\quad (T-8r)\times\alpha\times\Bigg(\frac{\sigma_m}{\sqrt{2\pi}\sqrt{1+2\gamma\frac{\psi^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}} + \frac{4\sigma_m}{\sqrt{2\pi}\sqrt{1+\gamma\frac{\psi^2}{\phi^2}+\rho\frac{\eta^2}{\phi^2}+\frac{1}{2r}\frac{1}{\phi^2}}}\Bigg) \tag{A.28}$$

The r which maximizes the above equation can be numerically determined.

Theorem 6: Optimal number of replicates for 2_{IV}^{7-2} design

The confounding structure for 2^{7-2} design is not same for all main effects. There are two types of confounding structure: (1) Three main effects are confounded with fourth order interaction effects and its estimates are from $N\big(|\Delta|,\big(\frac{\sigma_e^2}{8r}\big)\big)$ (2) Other main effects are confounded with third order interaction effects and its estimates are from $N\big(|\Delta|,\big(\rho\sigma_{3int}^2+\frac{\sigma_e^2}{8r}\big)\big)$. The expected improvement for 2^{7-2} design is given by

$$EI_{2_{IV}^{7-2}}(r) = \frac{3\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{8r}\frac{\sigma_e^2}{\sigma_m^2}}} + \frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\rho\frac{\sigma_{3int}^2}{\sigma_m^2}+\frac{1}{8r}\frac{\sigma_e^2}{\sigma_m^2}}} \tag{A.29}$$

Let $\frac{\sigma_m}{\sigma_e}=\phi$, $\frac{\sigma_{2int}}{\sigma_e}=\psi$ and $\frac{\sigma_{3int}}{\sigma_e}=\eta$

$$= \frac{3\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{8r}\frac{1}{\phi^2}}} + \frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\rho\frac{\eta^2}{\phi^2}+\frac{1}{8r}\frac{1}{\phi^2}}} \tag{A.30}$$

The optimal number of replicates for 2^{7-2} design is

$$\underset{r}{\text{Max}} \quad (T-32r)\times\left(\frac{3\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\frac{1}{8r}\frac{1}{\phi^2}}}+\frac{4\alpha\sigma_m}{\sqrt{2\pi}\sqrt{1+\rho\frac{\eta^2}{\phi^2}+\frac{1}{8r}\frac{1}{\phi^2}}}\right) \tag{A.31}$$

The r which maximizes the above equation can be numerically determined.

References

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p.

Berndtson, W. E. (1991). A simple, rapid and reliable method for selecting or assessing the number of replicates for animal experiments. *Journal of animal science*, 69(1):67–76.

Box, G. and Hunter, J. (1986). An analysis of unreplicated fractional factorials. *Technometrics*, 28(28):11–18.

Box, G. E. (1957). Evolutionary operation: A method for increasing industrial productivity. *Applied Statistics*, pages 81–101.

Fraser, D. A. and Guttman, I. (1956). Tolerance regions. *The Annals of Mathematical Statistics*, pages 162–179.

Frey, D. D. and Li, X. (2008). Using hierarchical probability models to evaluate robust parameter design methods. *Journal of Quality Technology*, 40(1):59.

Frey, D. D. and Wang, H. (2006). Adaptive one-factor-at-a-time experimentation and expected value of improvement. *Technometrics*, 48(3):418–431.

Gonzalez-Zugasti, J. P., Otto, K. N., and Baker, J. D. (2000). A method for architecting product platforms. *Research in engineering design*, 12(2):61–72.

Hamada, W. and Wu, C. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24:130–137.

İç, Y. T. (2016). Development of a new multi-criteria optimization method for engineering design problems. *Research in Engineering Design*, 27(4):413–436.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260.

Joseph, V. R. (2006). A bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48(2):219–229.

Katsikopoulos, K. V. (2009). Coherence and correspondence in engineering design: informing the conversation and connecting with judgment and decision-making research. *Judgment and Decision Making*, 4(2):147.

Li, X., Sudarsanam, N., and Frey, D. (2006). Regularities in data from factorial experiments. *Complexity*, 11(5):32–45.

Machin, D., Campbell, M. J., Tan, S.-B., and Tan, S.-H. (2011). Sample size tables for clinical studies.

Martin, M. V. and Ishii, K. (1997). Design for variety: development of complexity indices and design charts. In *Proceedings of*, pages 14–17.

Montgomery, D. C. (2008). *Design and analysis of experiments*. John Wiley & Sons.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, 17(8):857–872.

Orsborn, S., Boatwright, P., and Cagan, J. (2008). Identifying product shape relationships using principal component analysis. *Research in Engineering Design*, 18(4):163–180.

Osio, I. G. and Amon, C. H. (1996). An engineering design methodology with multistage bayesian surrogates and optimal sampling. *Research in Engineering Design*, 8(4):189–206.

Pham-Gia, T. and Turkkan, N. (1992). Sample size determination in bayesian analysis. *The Statistician*, pages 389–397.

Reich, Y. (2010). My method is better! *Research in Engineering Design*, 21(3):137–142.

Schlaifer, R. and Raiffa, H. (1961). Applied statistical decision theory.

Schmidt, S. R. and Launsby, R. G. (1989). *Understanding industrial designed experiments*. Air Academy Press.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., and Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2):322.

Simon, R. (1989). Optimal two-stage designs for phase ii clinical trials. *Controlled clinical trials*, 10(1):1–10.

Simon, R., Wittes, R., and Ellenberg, S. (1985). Randomized phase ii clinical trials. *Cancer treatment reports*, 69(12):1375–1381.

Soare, M. (2015). *Sequential Resource Allocation in Linear Stochastic Bandits*. PhD thesis, Université Lille 1-Sciences et Technologies.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16(3):243–258.

Sudarsanam, N. and Ravindran, B. (2017). Using linear stochastic bandits to extend traditional offline designed experiments to online settings. *Computers & Industrial Engineering*.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Tan, S.-B. and Machin, D. (2002). Bayesian two-stage designs for phase ii clinical trials. *Statistics in medicine*, 21(14):1991–2012.

Whitney, D. E. (1993). Nippondenso co. ltd: A case study of strategic product design. *Research in Engineering Design*, 5(1):1–20.

Willan, A. R. and Pinto, E. M. (2005). The value of information and optimal clinical trial design. *Statistics in medicine*, 24(12):1791–1806.