Conducting Non-adaptive Experiments in a Live Setting: A Bayesian Approach to Determining Optimal Sample Size

As Published: 10.1115/1.4045603

Publisher: ASME International

Persistent URL: https://hdl.handle.net/1721.1/136664

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Conducting Non-adaptive Experiments in a Live Setting: A Bayesian Approach to Determining Optimal Sample Size

Nandan Sudarsanam
Department of Management Studies,
Robert Bosch Center for Data Science and
AI (RBCDSAI),
IIT Madras,
Chennai, India
e-mail: nandan@iitm.ac.in

Ramya Chandran
Department of Management Studies,
IIT Madras,
Chennai, India
e-mail: ms16d021@smail.iitm.ac.in

Daniel D. Frey
Department of Mechanical Engineering,
MIT,
Cambridge, MA
e-mail: danfrey@mit.edu

This research studies the use of predetermined experimental plans in a live setting with a finite implementation horizon. In this context, we seek to determine the optimal experimental budget in different environments using a Bayesian framework. We derive theoretical results on the optimal allocation of resources to treatments with the objective of minimizing cumulative regret, a metric commonly used in online statistical learning. Our base case studies a setting with two treatments assuming Gaussian priors for the treatment means and noise distributions. We extend our study through analytical and semi-analytical techniques which explore worst-case bounds, the presence of unequal prior distributions, and the generalization to k *treatments. We determine theoretical limits for the experimental budget across all possible scenarios. The optimal level of experimentation that is recommended by this study varies extensively and depends on the experimental environment as well as the number of available units. This highlights the importance of such an approach which incorporates these factors to determine the budget.* [DOI: 10.1115/1.4045603]

Keywords: decision theory, design of experiments, design optimization, design theory and methodology, machine learning, uncertainty modeling

1 Introduction

Experimentation is an important tool for gaining knowledge about a system, process, or product. There have been several contributions from the design engineering community toward the betterment and usability of experimentation in mechanical systems. These contributions can be broadly classified into two groups based on the approach they adopt. The first, and more popular one, looks at traditional non-adaptive, multi-factor design of experiments (DOEs) to solve problems in multiple domains (examples include Refs. [1,2]). Other efforts have attempted to use an adaptive strategy, in which the experimentation is carried out in a sequential fashion [3–5]. In both the approaches, experimentation is considered to be carried out in an offline setting. This indicates a scenario where the output from the experiments is sacrificed toward the goal of future improvement. By contrast, the field of reinforcement learning in computer science addresses the problem of online learning through the multi-armed bandit (MAB) problem [6] (a survey can be found in Ref. [7]). Here, an agent sequentially chooses treatments with a goal of performing as best as possible in a live environment, cumulatively, over a given horizon of time or resources.

Motivated by various real-world examples that span multiple domains, this study seeks to address an experimentation problem that essentially lies in the intersection of these two well-studied fields: randomized experimental designs studied in statistics and the multi-armed bandit framework from reinforcement learning. In randomized experimentation, the typical setup requires an upfront design and commitment of experimental resources. The DOE approach is suited for an offline, learning-phase, which can support a parallel deployment of resources since it is non-adaptive (classic examples are in agriculture, product design, and manufacturing). Whereas, the MAB framework, which is intended to be applied on live systems, typically proposes an adaptive algorithm, with almost instantaneous feedback from experimenting on one-resource-at-a-time (typical examples are on online advertisements and recommender systems). Our research considers cases that are a mix of these two scenarios. In our scenarios, the environment requires the commitment and planning of a one-shot experiment, but simultaneously needs to in a live setting. Specifically, the environment we study is characterized by three specific features, an online environment, finite horizon or scope, and the need for preplanned, non-adaptive designs. In this context, the key research question we answer is in determining the upfront experimental commitment of resources. Our study seeks to construct a framework for experimentation in the presence of these conditions. Section 1.1 presents specific examples where such conditions could arise, and Sec. 1.2 provides the basic mathematical framework for the study.

1.1 Real-World Examples of the Environment. We illustrate three concrete examples where the conditions mentioned above could take place:

(1) A city receives $ 1,000,000 to be used in 1 year for improving 200 parks ($ 5000 per park). The funding is intended to increase the usage (footfalls) in the park. The Director of Parks and Recreation believes that there are two specific infrastructural interventions that cost $ 5000 each but is unsure on which one would increase footfalls. However, she does believe that she can run a pilot on some parks for 6 months, leaving sufficient time to observe the outcome and implement the superior treatment on the remaining parks. How many parks out of the 200 should she experiment on?

Contributed by the Design Automation Committee of ASME for publication in the Journal of Mechanical Design. Manuscript received June 9, 2019; final manuscript received November 1, 2019; published online December 5, 2019. Assoc. Editor: Nam H. Kim.

(2) The marketing division of an organization has identified 10,000 specific customers that are most likely to be interested in a discount on a particular product. They would like to send e-mails to these customers indicating a special offer. However, there are four possible e-mail designs that indicate the same offer (two different subject lines, and two different pictures in the body of the email, leading to four combinations). The marketing manager would like to know which of the four possible e-mail designs is more likely to attract customers. She would like to conduct an experiment by sending out each of the four designs to different customers covering a subset of the 10,000 customers. Following the recommended wait period of 15 days, she will analyze the customer response to the e-mails and intend to send the design showing most promise to the remaining customers. How many customers out of the 10,000 should be a part of the experiment?

(3) Building on an example from Ref. [8], a manufacturing plant has four computer numerical control machines that produce component parts to be assembled in an aircraft power unit. The engineer in charge knows that variability in the critical dimension of interest is affected by the feed rate. In a given order, there are a total of 2000 parts to be machined, and the time available for delivery would require all machines to run continuously, with at most one break for reconfiguration (as setup times are long and machines need to run in parallel this obviates any sequential learning). The engineer uses her domain knowledge to choose four slightly different feed rates to be applied to the four machines as an experiment. She is confident that all the parts can be shipped and are fit for use, but she would like to minimize the variability as much as possible. Following, her experiment she will apply the feed rate that appears to give the best result to the remaining three which will continue to reach the target of 2000 totally. How many parts should be machined with the different feed rates in the initial phase?

1.2 Problem Setup. In this section, we first consider a simplified version of the problem, where the experimenter is given a total set of units T and has to choose to allocate one of two treatments on these units. With no prior preferences, she chooses to conduct an experiment on n units of each of the two treatments. Based on the results, she picks the treatment with the highest mean and deploys that on the remaining $(T-2n)$ units. The critical design question that we solve is to determine what the sample size (n) should be for a given (T).

We use expected cumulative regret, a common metric used in the MAB framework to measure performance in the online environment. Here, the selection of a sub-optimal treatment on a unit, during the experiment or post-experiment, leads to some quantifiable regret. This regret aggregated across all the experimental units is cumulative regret. In a stochastic framework, where the experimental outcomes are also a result of irreducible noise, the cumulative regret for a given plan can be a random variable. We take the expected value of this variable. Mathematically, the expected cumulative regret is given by

$$E(CR) = n \times X + (T - 2n) \times (1 - \Pr(L^*(n))) \times X \qquad (1)$$

where X is the gap between the expected value of the optimal treatment and the suboptimal one. The term $\Pr(L^*(n))$ captures the probability that following the results of conducting $2n$ balanced experiments across both treatments we conclude that the optimal treatment is the best. The first part of the RHS in Eq. (1) corresponds to the idea that for we accrue a regret of X for n units due to the deployment of n suboptimal units during the experimentation phase. The latter half of Eq. (1) corresponds to the probabilistic loss that can be accrued if we fail to choose the optimal setting.

Intuitively, it can be reasoned that n should not be too high or too low for minimizing the cumulative regret. If n is extremely large (say $T/2$ units), then there would be no units left to apply our findings from the experiment. If n is too small, it is possible that we made an incorrect choice of the optimal treatment and subject the remaining units to this. This is captured in Fig. 1, which shows the intuition behind a trade-off associated with a degree of experimentation in the online framework. In this research, we mathematically formulate and derive the relationship between sample size and regret for various environments. We then seek to determine the optimal sample size (n^*) which minimizes the expected regret.

The rest of the paper is structured as follows: Sec. 2 covers the related literature, Sec. 3 covers the theory for the basic algorithm. This serves as a framework for all the possible extensions to the environment and algorithms covered in Sec. 4. We conclude with Sec. 5 and provide some directions for future work.

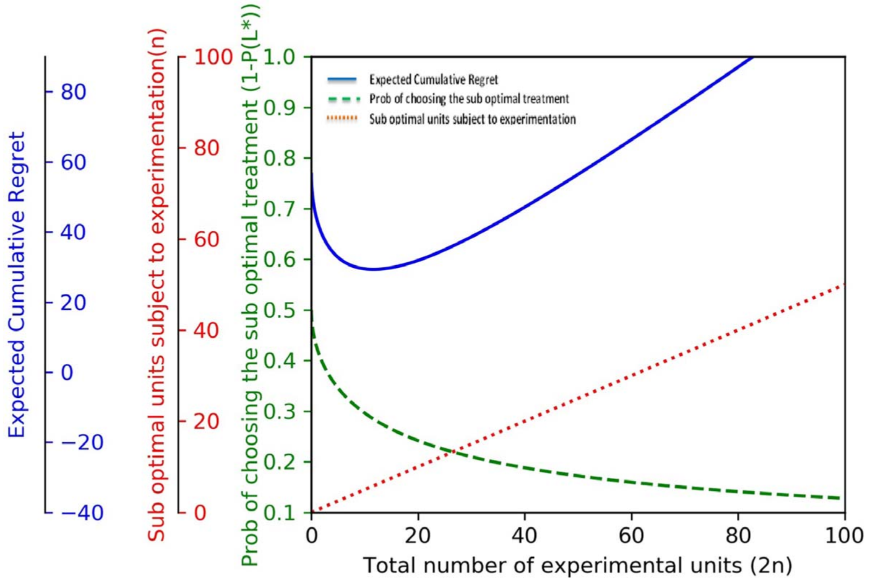

Fig. 1 Illustrative figure on the effect of changing sample sizes on expected cumulative regret

2 Related Work

Studies pertaining to experimentation can be broadly categorized as offline or online. Offline experimentation, or more specifically the determination of sample size of non-dynamic experimental plans, is further divided across Frequentist and Bayesian statistics. The online environment is discussed extensively by studies in Reinforcement Learning under the multi-armed bandit (MAB) framework. Our study considers non-dynamic plans in the online setting. In this section, we discuss contributions from both of these fields.

In frequentist statistics, the sample size is extensively discussed through operational characteristic (OC) curves. With traditional OC curves, the type II error is described as a mathematical function of sample size, type I error δ and σ (sometimes δ/σ). Various reviews of the well-established mathematical interrelationship between these variables in the context of hypothesis testing and estimation can be gathered from Ref. [9] or Ref. [10]. This is somewhat similar to what we are trying to achieve through Fig. 1 for the online setup. Our environment is, however, different in two ways. First, at the end of the experiments, the practitioner is required to choose a treatment (irrespective of statistical significance) to apply on the remaining units, therefore leaving no scope for Type II error. The important equivalent metric is the probability of choosing the superior/inferior treatment. This is captured in the Eq. (1). The other aspect of difference from OC curves is that there exists an optimal level of experimentation in our setting, as opposed to an asymptotic relationship between increased samples and reduction of Type II error. Similar to traditional OC curves, a greater sample size in our setting does lead to a higher probability of choosing the correct treatment. However, in the finite population online setting, a large sample allocation to experiments will leave fewer units to apply the experimental findings (this is often referred to as the exploration-exploitation dilemma in bandit studies).

Broadly, the idea that the cost of gathering a sample should be compared with its potential benefits that has been well-studied by using a Bayesian approach in the decision theory. We can see examples of this in Expected Value of Sample Information [11,12] and other studies that build on this idea [13]. The motivation for the Bayesian approach is that a single sample may not be sufficiently large to make an inference about the population parameter and also studies are rarely carried out in complete isolation. A Bayesian approach enables us to assume prior distributions for the parameters of interest and thereby explicitly compare the costs of experimenting against the perceived benefits. These have motivated studies which look at both static plans [9,10,14] as well as sequential ones [15–17] to optimize experimental decision-making. Our study adopts a similar idea. However, in our live setting, the cost of over-experimenting is the allocation of sub-optimal treatments to units as a part of the experiment, and the cost of under-experimenting is the increased probability of selecting sub-optimal treatments for the remaining units. These two costs are directly comparable, as opposed to the traditional setups where the decision-making required the practitioner to quantify external costs and benefits related to the experimental process. This broad idea is explored in the two-level full and fractional factorial designs by Ref. [18] using a metric of cumulative improvement.

Experimenting on live systems has been studied since as early as 1930s under the terminology of the sequential design of experiments [19–21]. This has been advanced extensively in the multi-armed bandit framework discussed in the reinforcement learning literature. Here, an agent is tasked with sequentially choosing a treatment (referred to as *arms*) for each trial, with the goal of maximizing the cumulative reward across t trials [6,22]. Since the agent does not have perfect knowledge of which treatment is superior, there is an inherent need to explore (experiment) across treatments. Also, since the horizon is finite (t trials) or discounted, there is a cost to conducting too many experiments. As discussed in Sec. 1, there are various settings where the bandit framework is not entirely suitable. These are broadly summarized as the need to parallelize the experimental effort and the concern that a fully sequential approach is harder to implement. The later concern comes about since the practitioner is unaware of the number of actual experiments that each alternative will be subject to, prior to experimentation. In this study, these observations motivate us to consider environments where a randomized experiment along with the sample sizes is fixed up front.

Similar algorithms and analyses from the bandit community have predated our work. Reference [23] have made a study on the Batched Bandit problems. However, their analysis differs from ours in that they provide only worst-case bounds (whereas this study minimizes expected cumulative regret), do not make any distributional assumptions on the gap (whereas this study models the gap as a random variable), and finally, their work seeks to provide a dynamic stopping point (whereas ours reflects a common real-world setting that requires a planned budget). Another set of bandit studies that allow for the parallel deployment of treatments are round-based algorithms [24]. Here, each treatment is tried once in a given round and criteria for the elimination of sub-optimal treatments enable the convergence to the optimal treatment asymptotically. Again, our work allows for multiple experimental units to be subject in parallel to the same treatment in a given round.

It is also noteworthy that many of the extensions considered in this study are inspired by similar extensions in the offline environment or MAB algorithms, although none of these directly talk about our problem context. For instance, worst-case scenarios have been looked at by few studies in statistical experiments [25,26] and is almost a standard practice in bandit studies. Similarly, the extension of sample size calculations to multiple treatments are well studied through the use of the ANOVA, and the cases of using different distributions are also common [27]. Finally, multiple rounds or phases of pilots have also been considered by various authors in frequentist and Bayesian setups [28–30].

3 Theory: Analytical Results for Two-Treatments With Gaussian Assumptions

In this section, we introduce a theoretical analysis for the simple case where there are two treatments. We are tasked with picking one of the two treatments for each of the T available units. We have an added constraint that our commitment of treatment to resources can only be done in two phases. It can be easily shown that with no prior preferences between treatments, the first phase would ideally allocate resources in a balanced fashion between the two treatments. Similarly, based on the results of the first phase, the allocation of all resources to the ostensibly superior treatment would be an optimal strategy for the second phase. We, therefore, undertake the key question of how to divide resources across the two phases. Following the structure shown in Eq. (1), it is clear that the optimal number of experiments (n^*) is a function of the gap (X) between optimal and sub-optimal treatment, the total number of units (T), and probability of identifying the optimal treatment after the experiments (which are in turn a function of X and the noise in the system). Of these, the exact knowledge, or even estimates, of the value of X could be an unrealistic requirement. Essentially, we are conducting experiments to determine which treatment has a superior mean, and our assumptions on X imply that the practitioner knows the quantum of difference between the two means. While these assumptions are standard in typical OC curve calculations, in the offline environment an increase in X translates to a monotonic decrease in sample size for fixed Type II error. This helps to facilitate the understanding that a given sample size translates to a given Type II error for a given minimum value of X (larger values will only translate to lower Type II errors). However, in the online setting, a given X translates to an optimal recommendation of sample size and any differences in value, larger or smaller, will necessarily translate to suboptimal behavior. A more flexible and feasible assumption would be to model X as a random variable, rather than as a fixed value. This could capture the uncertainty in the prior belief of the likely magnitude of the difference in means. For instance, take an advertising firm which is proposing to experiment with two advertising initiatives to increase sales (on a fixed target audience or cities), it is unlikely that they know which treatment is superior (which is the reason for the experiment) or the magnitude of the difference in the mean effect of the initiatives. However, it is possible that prior experience with such initiatives, in general, could be used to inform them of the likely distribution of effect seen when such initiatives are taken. This can then be used to derive the distribution of the difference in means of the two randomly selected treatments. Here, we assume that the true mean of each treatment, μ_1 and μ_2, has a prior Gaussian distribution $N(\mu, (\sigma_m/\sqrt{2})^2)$. Then, the distribution of $\mu_1 - \mu_2$, referred to as $X \sim N(0, \sigma_m^2)$, or in other words, the magnitude of the difference between the two treatments is distributed as a half-normal, a special case of the folded normal. We also assume a setup where the error/noise per experimental trial follows the Normal distribution $N(0, \sigma_e^2)$. The probability of choosing the superior treatment is then formulated by representing the distribution of $|X|$ as a half-normal distribution and computing the probability that the estimate represented as a conditional random variable $f_{Y|X}(y) \sim N(|x|, 2\sigma_e^2/n)$, takes on a value greater than 0. If L^* is the superior treatment, then the formulation below captures the probability of choosing L^*:

$$\Pr(L^*) = \int_0^\infty \int_0^\infty \frac{1}{\frac{\sigma_e}{\sqrt{n}} 2\sqrt{\pi}} exp\left(-\frac{(y-x)^2}{\frac{4\sigma_e^2}{n}}\right) dy \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp\left(-\frac{x^2}{2\sigma_m^2}\right) dx \tag{2}$$

The expected cumulative regret ($E(CR_{2n})$) is given by

$$\begin{aligned} E(CR_{2n}) = & \int_0^\infty \int_0^\infty n \times x \times \frac{1}{\frac{\sigma_e}{\sqrt{n}} 2\sqrt{\pi}} exp\left(-\frac{(y-x)^2}{\frac{4\sigma_e^2}{n}}\right) \\ & \times dy \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp\left(-\frac{x^2}{2\sigma_m^2}\right) dx \\ & + \int_0^\infty \int_{-\infty}^0 (T-n) \times x \times \frac{1}{\frac{\sigma_e}{\sqrt{n}} 2\sqrt{\pi}} exp\left(-\frac{(y-x)^2}{\frac{4\sigma_e^2}{n}}\right) \\ & \times dy \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp\left(-\frac{x^2}{2\sigma_m^2}\right) dx \end{aligned} \tag{3}$$

On solving we get,

$$E(CR_{2n}) = \frac{\sigma_m n\left(\frac{\sigma_m}{\sqrt{\frac{2\sigma_e^2}{n}+\sigma_m^2}}+1\right)}{\sqrt{2\pi}} + \frac{\sigma_m(T-n)\left(1-\sigma_m\sqrt{\frac{n}{2\sigma_e^2+n\sigma_m^2}}\right)}{\sqrt{2\pi}} \tag{4}$$

It could be convenient to express this regret in percentage terms. We achieve that by defining the *maximum-expected* regret, which is the regret experienced when the suboptimal treatment is applied over the entire horizon T. However, the gap between the optimal and suboptimal treatment is a random variable. Therefore, we take the expected value of the gap between the optimal and suboptimal treatment, which can be shown to be $\sqrt{2/\pi}\sigma_m$ (this is essentially the mean of the half-normal distribution) and assume that this regret is experienced over the entire horizon. Therefore, the normalized regret for this case is

$$NECR_{2n} = \frac{\frac{\sigma_m\left(n\left(\frac{\sigma_m}{\sqrt{\frac{2\sigma_e^2}{n}+\sigma_m^2}}+1\right)+(T-n)\left(1-\sigma_m\sqrt{\frac{n}{2\sigma_e^2+n\sigma_m^2}}\right)\right)}{\sqrt{2\pi}}}{T\sqrt{\frac{2}{\pi}}\sigma_m} \tag{5}$$

Let $\Omega = \sigma_m/\sigma_e$, then, the above equation becomes

$$NECR_{2n} = \frac{1+\left(\frac{2n}{T}-1\right)\left(\sigma_m\sqrt{\frac{n}{2+n\Omega^2}}\right)}{2} \tag{6}$$

From Eq. (6), it can be seen that normalized expected cumulative regret is a function of n, T, and $\Omega = \sigma_m/\sigma_e$. Figure 2 illustrates this for different values of n, while fixing $T = 100$ and analyzing three ratios of Ω.

There are various insights that can be gleaned from these results. As expected, there is an optimal level of experimentation that minimizes regret, any more or lesser experimentation is suboptimal. However, it is clear that the cost of under-experimenting is worse than over-experimenting. Also, a higher value of Ω (or σ_m/σ_e) leads to lower normalized regret and requires a lower commitment to experimental resources. In other words, the higher signal-to-noise ratio is generally a favorable environment and requires a lesser commitment to exploration for the same certainty. While Fig. 2 computes the regret for various levels of experimentation, a practitioner could be interested in directly determining the optimal level of experimentation. As a next step, we determine the n^* which minimizes Eq. (6). This gives us the optimal sample size.

$$\min_n \frac{\sigma_m\left(n\left(\frac{\sigma_m}{\sqrt{\frac{2\sigma_e^2}{n}+\sigma_m^2}}+1\right)+(T-n)\left(1-\sigma_m\sqrt{\frac{n}{2\sigma_e^2+n\sigma_m^2}}\right)\right)}{\sqrt{2\pi}} \tag{7}$$

To find the maxima, we solve the $\partial NECR_{2n}/\partial n = 0$, which results in

$$n^* = \frac{-3\sigma_e^2+\sqrt{9\sigma_e^4+2T\sigma_e^2\sigma_m^2}}{2\sigma_m^2} \tag{8}$$

Let $\Omega = \sigma_m/\sigma_e$, then the optimal sample size becomes

$$n^* = \frac{-3+\sqrt{9+2T\Omega^2}}{2\Omega^2} \tag{9}$$

The use of the Gaussian assumption on X results in a closed-form expression for the optimal sample size as a function of only T and Ω. Figure 3 shows the sensitivity of the optimal size for different values of Ω. This is illustrated for different values of $T = 100$, 1000, and 10,000. Figure 3 shows that for all values of T, an increase in Ω translates to a decrease in the required sample size for optimal performance. In absolute terms, the optimal level of experimentation (n^*) increases with an increase in T (since there are more units at stake when we choose to commit). However, it is noteworthy that the optimal n^* as a percentage of T becomes smaller with an increase in T, for all values of Ω. For instance, the optimal sample size for $T = 100$ at $\Omega = 1$ is 5.72, resulting in a ratio of approximately 6%, whereas, when $T = 10{,}000$, the optimal sample size is 69.2, resulting in a ratio of 0.7%.

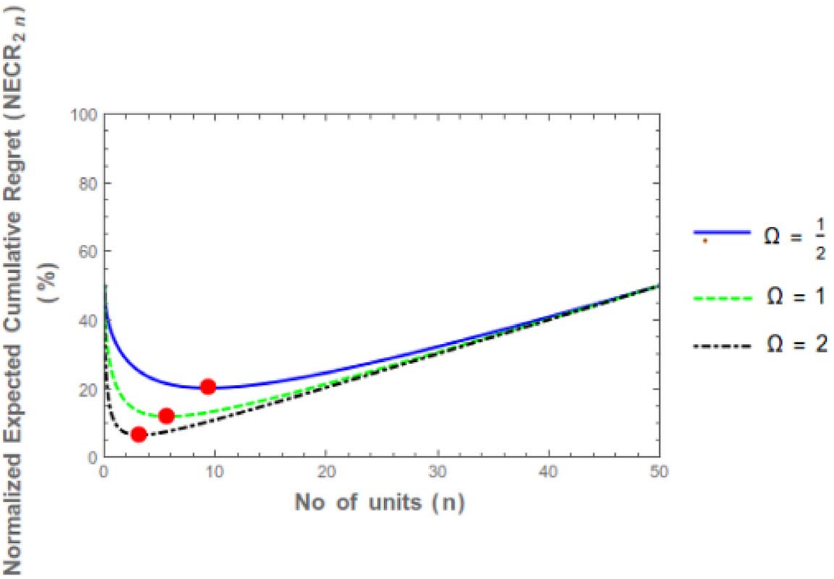

Fig. 2 Normalized expected cumulative regret over different sample sizes

4 Extensions to the Base Model

In this section, we present three extensions to the base model which increases the scope for application of this study. Section 4.1 explores the scenario where a practitioner is not interested in minimizing the expected value of cumulative regret, but would instead like to implement a risk-averse approach and would therefore minimize the worst-case scenario. Section 4.2 explores the scenario where the practitioners priors of the treatments are unequal. This could come about due to subjective beliefs, analysis of historical data, or even an additional pilot phase of experiments that were conducted prior to this analysis. The key emphasis in this section is in understanding the requirement for conducting unbalanced experiments in the online setting. Finally, Sec. 4.3 extends the analysis to cases where three or more treatments of the same variable might be present. This is a common occurrence in many real-world problems, and we derive analytical solutions for the optimal sample size.

4.1 Analysis of the Worst-Case Scenario. The base case analysis presented in Sec. 3 presents an optimal sample size based on averaging outcomes and therefore uses the expected value of cumulative regret. The experimenter might, in many instances, not be interested in minimizing the regret on average, and could instead be looking at reacting to the worst-case scenario. This approach is typical in bandit studies that look at worst-case regret bounds. We use the Gaussian setup to quantify this scenario represented by various percentiles of the entire distribution associated with cumulative regret. We present results for the case of the 90th, 95th, and 99th percentiles/bounds when $T = 100$ and $\Omega = 1$ (we can take any pair of values for σ_m and σ_e which satisfy this ratio). We adopt the following steps to find an optimal sample size through worst-case analysis: (i) model the entire distribution of total cumulative regret variable, (ii) construct an upper bound for it by fixing a percentile, and (iii) find a sample size that optimizes this bound.

We follow the same assumptions as we made for average-case analysis. The true mean of each treatment, μ_1 and μ_2, has a prior Gaussian distribution $N(\mu, (\sigma_m/\sqrt{2})^2)$. Then, the distribution of $\mu_1 - \mu_2$, referred to as $X \sim N(0, \sigma_m^2)$. Let $F_{C|N}(c)$ be the CDF of total cumulative regret for a given n (each treatment is subject to

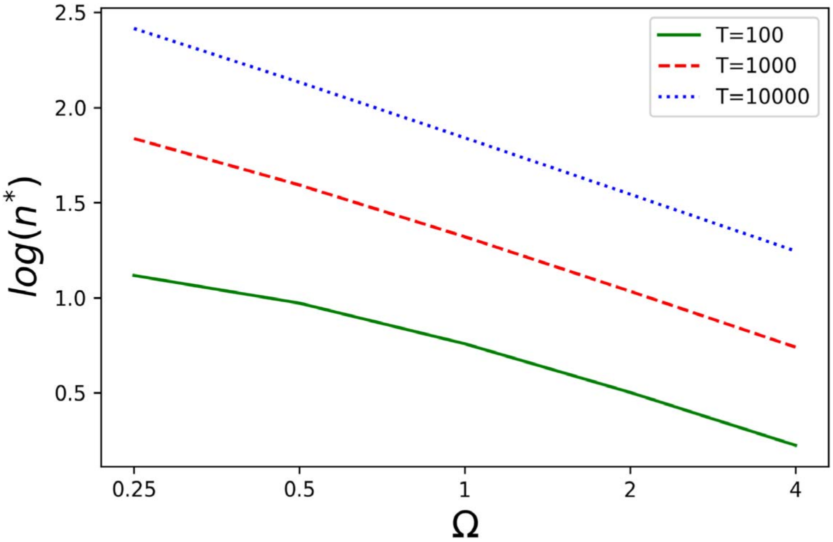

Fig. 3 Optimal sample size over different Ω

n trials therefore $2n$ resources are used for experimentation). This is essentially the $P(C \leq c|N=n)$. The inverse of the cumulative distribution function can be written as $F^{-1}_{C|N}(p)=c$. This tells us we can expect to see a cumulative regret below the value of c with probability p (for any given p). The $n*$ which minimizes this function gives the optimal sample size for the given worst-case bound characterized by p. In other words, $\min_{n} F^{-1}_{C|N}(p)$. However, without a closed form for $F^{-1}_{C|N}(p)$, we numerically determine c in by equating $p=F_{C|N}(c)$ for given bound values p for various values of n. We determine the optimal n through this method.

The CDF of $F_{C|N}(c)$ can be constructed by breaking down the probability into two parts.

$$F_{C|N}(c) = P(C \leq c \text{ and Picks the optimal treatment}|N=n)$$
$$P(C \leq c|N=n \text{ and Picks the sub-optimal treatment}) \tag{10}$$

We can state that $P(C \leq c|N=n$ and Picks the optimal treatment) occurs if X takes on a value such that the total regret $n \times x \leq c$. This is because the only regret we experience when we pick the optimal treatment is the n resources that were committed to the sub-optimal treatment during the experimental phase. Rearranging the terms, we get the inequality that $X \leq c/n$. Similarly, we can state that $P(C \leq c|N=n$ and Picks the sub-optimal treatment) occurs if X takes on a value such that the total regret $n \times x + (T-2n) \times x \leq c$. This is because, in addition to the regret from the n units during the suboptimal phase, we deploy $(T-2n)$ units sub-optimally. Rearranging the terms, we get the inequality that $x \leq c/(T-n)$. We use these inequalities to determine c for a given p, T, σ_m, σ_e for various values of n through the formulation below:

$$p = \int_0^{c/n}\int_0^{\infty} \frac{1}{\frac{2\sigma_e}{\sqrt{n}}\sqrt{\pi}} exp\left(-\frac{(y-x)^2}{\frac{4\sigma_e^2}{n}}\right) dy \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp\left(-\frac{x^2}{2\sigma_m^2}\right) dx$$
$$+ \int_0^{c/(T-n)}\int_{-\infty}^{0} \frac{1}{\frac{2\sigma_e}{\sqrt{n}}\sqrt{\pi}} exp\left(-\frac{(y-x)^2}{\frac{8\sigma_e^2}{n}}\right) dy \frac{\sqrt{2}}{\sigma_m\sqrt{\pi}} exp\left(-\frac{x^2}{2\sigma_m^2}\right) dx \tag{11}$$

Similar to Sec. 3, we can normalize this regret by dividing it by the *maximum-expected* regret which can be defined by $T\sqrt{2/\pi}\sigma_m$. As explained in Sec. 3, we assume that the suboptimal treatment is selected over the entire trial horizon T, and the average gap between the optimal and sub-optimal treatment is $\sqrt{2/\pi}\sigma_m$. Figure 4 shows the 90th, 95th, and 99th percentiles of normalized cumulative regret and compares this to the expected value for $T=100$ and $\Omega=1$. There are three interesting findings from Fig. 4 which we discuss here. First, the normalized worst-case regret can and does exceed a 100%, and this is because our definition is based on *maximum-expected* regret (see Sec. 3), which is a normalization over the maximum possible regret for the average gap between optimal and sub-optimal treatments. The true distribution for the cumulative regret would also account for the uncertainty in the gap between the treatments, and therefore, a worst-case scenario could be worse than the *maximum-expected* regret. The second finding is that as p increases (the degree of pessimism), this leads to an overall increase in cumulative regret and motivates more experimentation. Informally stated, if one expects to be unlucky in our guess of the optimal treatment, we prefer to hedge for this by becoming more conservative, and this is achieved by experimenting more. Finally, the third finding we discuss is that the expected value of cumulative regret is similar to the 90% worst-case scenarios plotted in the graph. In fact, the use of an 85% bound would result in lower regret at optima than the expected value. The reason for this is the nature of the distribution of cumulative regret. The cumulative regret is bimodal in nature, corresponding to the two modes of (i) guessing the optimal treatment, in which case the regret is confined to the experimental phase, or (ii) failing to guess the optimal treatment, in which case the regret is experienced over the remaining units. Under conditions close to the optimal level of experimentation, the majority of the density lies in the lower mode, with the remaining (minor) portion of the density at a much higher value. This has the effect of the long, heavy tail resulting in the expected value being close to the 85% percentile mark in the density function. In other words, the use of expected value inherently serves to prescribe a risk-averse behavior.

4.2 Unequal Priors on Treatment Means.

In Sec. 3, we determine the sample size for a balanced experiment under the assumption that the true mean of each treatment, μ_1 and μ_2, is randomly sampled from a Gaussian distribution $N(K, (\sigma_m/\sqrt{2})^2)$, and therefore the gap between means $X \sim N(0, \sigma_m^2)$. In this section, we explore the case where X has a non-zero mean. This could arise from historical data or experimenter's priors that favor one treatment over the other, probabilistically. While it would be fairly

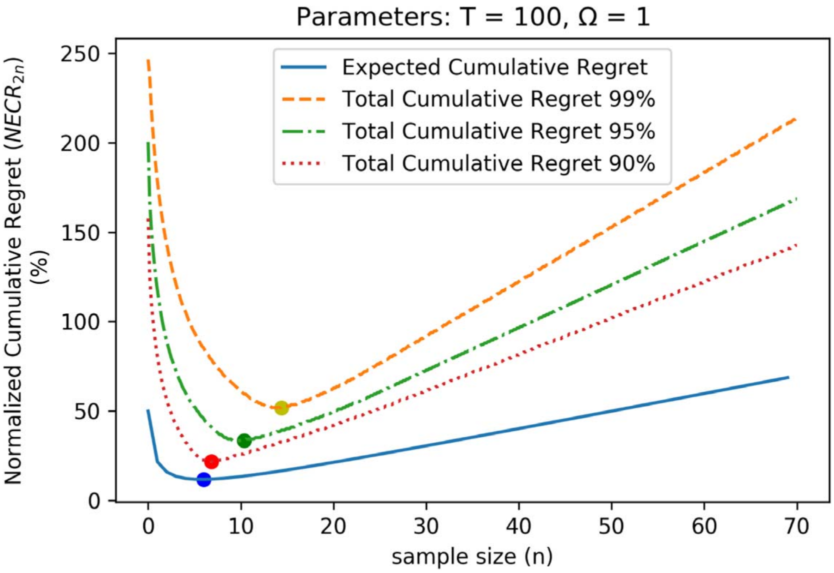

Fig. 4 Normalized cumulative regret over different sample sizes

straightforward to show that a non-zero mean for X would still favor balanced experimentation in the offline environment (where the goal is the only maximization of information), this setting has some interesting implications for experimentation in the online environment. Here, we no longer favor balanced experimentation of the treatments. Intuitively, there is now an added trade-off even in the experimental phase, where favoring the ostensibly superior treatment (as suggested by the prior distributions) will probabilistically lead to lower regret, whereas continuing to stay as close to a balanced design will provide the highest scope for identifying the truly superior treatment (which can be exploited in the next phase).

In line with terminology and formulation from Sec. 3, here we assume that the prior on mean of treatment A is $f_A(a) \sim N(\mu_A, \sigma_m^2/2)$ and treatment B is $f_B(b) \sim N(\mu_B, \sigma_m^2/2)$ if $\mu_m = \mu_A - \mu_B$ then the gap $X = A - B$ is given by $f_X(x) \sim N(\mu_m, \sigma_m^2)$. If the noise in the system follows $N(0, \sigma_e^2)$. After n_1 samples of A and n_2 samples of B, the conditional distribution of the estimate is $f_{Y|X}(y) \sim N(x, \sigma_e^2/n_1 + \sigma_e^2/n_2)$. When $\mu_m = 0$ (the special case studied in Sec. 3), a decision to choose the treatment with the higher sample mean is in line with both a frequentist and Bayesian decision-making process. This is the analysis we carried out in Sec. 3. However, when $\mu_m \neq 0$, a Bayesian decision-making framework would use the samples to update the posterior means and choose the treatment with the higher posterior mean.

Using the Gaussian assumptions, the conjugate posterior of the mean is

$$\mu_{post} = \left(\frac{1}{\dfrac{1}{\sigma_m^2} + \dfrac{1}{\dfrac{\sigma_e^2}{n_1} + \dfrac{\sigma_e^2}{n_2}}} \right) \left(\frac{\mu_m}{\sigma_m^2} + \frac{y}{\dfrac{\sigma_e^2}{n_1} + \dfrac{\sigma_e^2}{n_2}} \right) \tag{12}$$

From the posterior mean, we can make a claim that if y is greater than $(-(\sigma_e^2/n_1 + \sigma_e^2/n_2)\mu_m)/\sigma_m^2$, then our μ_{post} will be greater than zero leading to a decision to pick treatment A, which was subject to n_1 samples in the experimental phase.

The expected cumulative regret can be written as

$$\begin{aligned} E(CR_{2n}) = & \int_0^\infty n_2 \times x f_X(x) \int_{(-(\sigma_e^2/n_1+\sigma_e^2/n_2)\mu_m)/\sigma_m^2}^{\infty} f_{Y|X}(y)\, dy\, dx \\ & + \int_{-\infty}^{0} (T - n_2) \times -x f_X(x) \int_{(-(\sigma_e^2/n_1+\sigma_e^2/n_2)\mu_m)/\sigma_m^2}^{\infty} f_{Y|X}(y)\, dy\, dx \\ & + \int_{-\infty}^{0} n_1 \times -x f_X(x) \int_{-\infty}^{(-(\sigma_e^2/n_1+\sigma_e^2/n_2)\mu_m)/\sigma_m^2} f_{Y|X}(y)\, dy\, dx \\ & + \int_{0}^{\infty} (T - n_1) \times x f_X(x) \int_{-\infty}^{(-(\sigma_e^2/n_1+\sigma_e^2/n_2)\mu_m)/\sigma_m^2} f_{Y|X}(y)\, dy\, dx \end{aligned} \tag{13}$$

Where, as previously mentioned, $f_X(x) \sim N(\mu_m, \sigma_m^2)$ and $f_{Y|X}(y) \sim N(x, \sigma_e^2/n_1 + \sigma_e^2/n_2)$. The pair $(n_1{*}, n_2{*})$ which minimizes the above equation gives the optimal sample allocation for the treatments. This can be numerically evaluated. We explore a few illustrative values to provide readers with the intuition of how an unequal prior motivates the experimenter to utilize more resources in experimenting on the favorable treatment, in the online context. We adopt the case where $T = 100$, $\sigma_m = 1$, and $\sigma_e = 2$. This is one of the settings explored in Sec. 3 and shown in Fig. 2. Also, it represents the more pathological environment, which requires more experimentation. For these values, we explore three different values of $\mu_m = 0$, $\mu_m = 0.4316$, and $\mu_m = 0.9678$. These values have been specifically selected such that they result in a mean and variance combination which translates to the probability that treatment A is better than treatment B by 0.5, 0.6667, and 0.8334, respectively. Contour plots in Figs. 5–7 show the results of the cumulative regret (contour lines) for different values of n_1 and n_2.

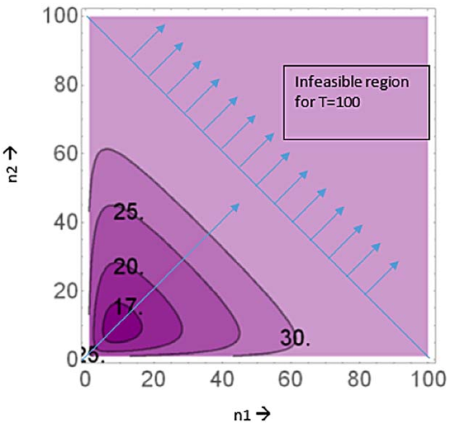

Fig. 5 ECR contour plot for $\mu_m = 0$

Figure 5 is a baseline case where both priors are equally likely. In essence, this would conform to the results of Sec. 3. While Sec. 3 assumes that the experiments need to be balanced for optimal performance, or in other words $n = n_1 = n_2$, this analysis confirms that the balanced trials result in the best performance. The 45° diagonal from the origin indicates the cases where $n_1 = n_2$ and in Fig. 5 we see that the optimal values are reached at $n_1 = n_2 = 9.5$ (which is inline with Fig. 2. It can also be seen that the balanced designs are also best possible allocation for any given budget (note that the 45° diagonal perpendicular to the one from the origin, connecting the positive and x and y-axes represents a contour line of a given budget for total experimentation or $n_1 + n_2 = constant$). Furthermore, this graph shows us the rate at which regret increases when experiments are conducted in an unbalanced way.

In other cases, we have priors that imply that one treatment mean is greater than the other ($\mu_m \neq 0$). When priors indicate this, in an online environment, a new trade-off takes place. On the one hand, a balanced experiment continues to have the steepest gain in determining which treatment is truly superior. This would contribute the most in minimizing regret during the post-experimental phase. By contrast, a commitment to the purportedly superior treatment provides benefits of minimizing regret during the experimental phase. Overall, this results in leaning toward more experiments on the preferred treatment. In Fig. 6, for instance, when $\mu_m = 0.4316$, corresponding to the likelihood that the first treatment is 66.66% likely to be the optimal one, we see that optimal allocation is seen when $n_1 = 13$ and $n_2 = 6$. As we keep increasing the likelihood that a treatment is better than the other we find that this results in minimal motivation to experiment on the inferior treatment. This is seen in Fig. 7 where the $\mu_m = 0.9678$, corresponding

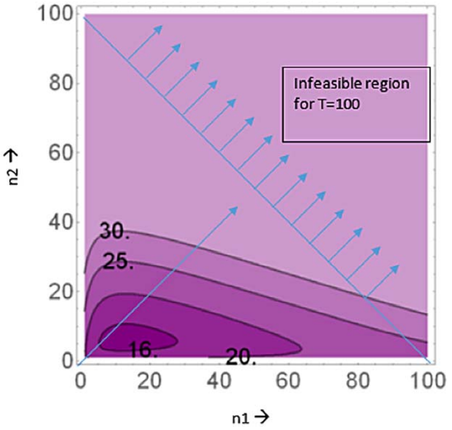

Fig. 6 ECR contour plot for $\mu_m = 0.4316$

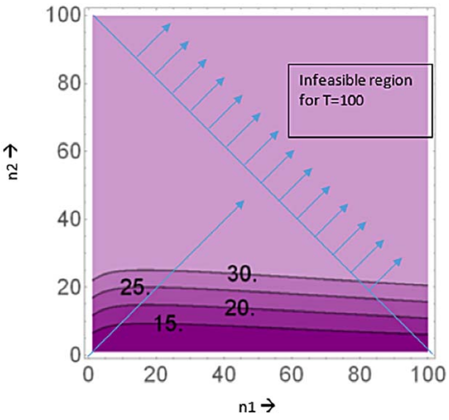

Fig. 7 ECR contour plot for $\mu_m = 0.9678$

to the probability that the mean of the first treatment is 83.33% more likely than the second one to be superior. Here, the optimal allocation is in expending no resources on the probabilistically inferior treatment (our constraints require 1 unit as a base). Interestingly, we are largely indifferent to the commitment of experimental resources to the ostensibly superior treatment for a given commitment of resources to the inferior one (the contour curves in Fig. 7 are almost horizontal). These two findings stem from the same phenomena. Given the priors in such an environment, exploration has little effect on changing the decision from the superior to the inferior treatment based on sample evidence (the relatively large μ_m would imply that $\bar{x}$ would have a lesser effect on μ_{post} in Eq. (12)). Therefore, low experimentation, with a high probability, will lead to expending all resources on the seemingly superior treatment. Similarly, experimenting blindly on the seemingly superior treatment has the same effect, and risk. This results in a flattish response to varying degrees of experimentation on the treatment which is theoretically expected to have a higher mean.

4.3 Analysis of Multiple Treatments.

In Sec. 3, we use expected cumulative regret as a performance measure that needs to be minimized, and we determine the optimal sample size for an experiment which achieves this objective. However, when there are more than two treatments or, stated another way, that the factor of interest has three or more levels, this conception proves to be challenging. In this section, we explore expected cumulative improvement ($E(CI_{kn})$), which is the exact complement of cumulative regret, with respect to average performance across all treatments. Essentially, expected cumulative regret is a measure of the average cumulative regret benchmarked against optimal performance (defined by always choosing the optimal treatment), the expected cumulative improvement captures the average improvement in performance benchmarked against a truly random selection of treatments. An experimental budget that minimizes expected cumulative regret would identically be maximizing expected cumulative improvement. Its use can also be seen in other studies such as Ref. [31]. The cumulative improvement for T when subjecting each of the k treatments to n replicates is defined as

$$CI_{kn} = (T - n \times k) \times \text{Expected Improvement} \tag{14}$$

where the expected improvement is defined as the average improvement in performance that can be expected from picking the treatment which resulted in the highest mean after $n \times k$ experiments. Understandably, this improvement is exploited only on the remaining $T - n \times k$ units and the first $n \times k$ have 0 improvement over the baseline of a random selection owing to the fact that this is a balanced experiment. We model each of k levels of a factor's mean response as being drawn from a normally distributed population with a mean of 0 and variance of $\sigma_m^2/2$.[1] We are interested in assessing the improvement in the outcome attained by estimating those factor levels through an experiment replicated n times with pure experimental error σ_e^2. We assume that $n \times k$ which is the total experimental budget is less than T.

The expected value of improvements in the results after the experimentation phase across different values of k are as follows: using the same terminology adopted in Sec. 3, similar to Eq. (3), for $k = 2$ levels the expected improvement can be written as

$$\begin{aligned} EI_{2n} = & \int_0^\infty \int_0^\infty \frac{x}{2} \times \frac{1}{\frac{\sigma_e}{\sqrt{n}} 2\sqrt{\pi}} exp\left(- \frac{(y-x)^2}{\frac{4\sigma_e^2}{n}} \right) \\ & \times dy \frac{\sqrt{2}}{\sigma_m \sqrt{\pi}} exp\left(- \frac{x^2}{2\sigma_m^2} \right) dx \\ & + \int_0^\infty \int_{-\infty}^0 \frac{x}{2} \times x \times \frac{1}{\frac{\sigma_e}{\sqrt{n}} 2\sqrt{\pi}} exp\left(- \frac{(y-x)^2}{\frac{4\sigma_e^2}{n}} \right) \\ & \times dy \frac{\sqrt{2}}{\sigma_m \sqrt{\pi}} exp\left(- \frac{x^2}{2\sigma_m^2} \right) dx \end{aligned} \tag{15}$$

$$EI_{2n} = \frac{1}{\sqrt{\pi}} \frac{\sigma_m^2}{2\sqrt{\frac{\sigma_e^2}{n} + \frac{\sigma_m^2}{2}}} \tag{16}$$

Therefore, the cumulative improvement for two treatments case is

$$E(CI_{2n}) = (T - 2n) \times \frac{1}{\sqrt{\pi}} \frac{\sigma_m^2}{2\sqrt{\frac{\sigma_e^2}{n} + \frac{\sigma_m^2}{2}}} \tag{17}$$

In order to enable the generalization of these results to any number of treatments, we use the well-established relationship in order statistics that if X_i, $i \in_1 \dots k$ are independent and normally distributed then $E(\max_i X_i) = \mu + \sigma \int_{-\infty}^{\infty} t(d/dt)(\Phi(t)^k) dt$. We now rewrite $E(CI_{2n})$ in a generalizable form

$$E(CI_{2n}) = (T - 2n) \times \frac{\sigma_m^2}{2\sqrt{\frac{\sigma_e^2}{n} + \frac{\sigma_m^2}{2}}} \times \int_{-\infty}^{\infty} t \frac{d}{dt} \Phi(t)^2 \, dt \tag{18}$$

where it can be shown that $\int_{-\infty}^{\infty} t(d/dt)\Phi(t)^2 \, dt = 1/\sqrt{\pi}$.

This allows us to extend our results for k treatments.

$$E(CI_{kn}) = (T - kn) \times \frac{\sigma_m^2}{2\sqrt{\frac{\sigma_e^2}{n} + \frac{\sigma_m^2}{2}}} \times \int_{-\infty}^{\infty} t \frac{d}{dt} \Phi(t)^k \, dt \tag{19}$$

While our primary objective is to determine the optimal n for any given k, we first discuss some interesting insights that can be seen in Eq. (19). The values of $\int_{-\infty}^{\infty} t(d/dt)(\Phi(t)^k) \, dt$ for various values of k can be determined analytically for $k = 2, 3, \dots, 6$. (summarized in Ref. [32], refer to Table 1).

A plot of these values (Fig. 8), which essentially captures the benefit of having more treatments, shows that the relationship is monotonic but is convex (steep initially but plateaus out). By contrast, the experimental effort increases linearly as k increases (the $T - nk$ term in Eq. (19)). This indicates that for a given value of n, increases in k could initially lead to an increase in cumulative improvement (if the curve of $\int_{-\infty}^{\infty} t(d/dt)(\Phi(t)^k) \, dt$ is steeper than the linear drop from $T - nk$), but will certainly lead to a point where increase in k is detrimental. However, if n* (the optimal n)

[1] To be consistent with assumptions in Secs. 3, 4.1, and 4.2, the treatment or factor's means are sampled from $N(\mu, (\sigma_m/\sqrt{2})^2)$.

Table 1 Analytical results of $\int_{-\infty}^{\infty} t(d/dt)(\Phi(t)^k)\,dt$ for various values of k from Ref. [32]

k	$\int_{-\infty}^{\infty} t\frac{d}{dt}\Phi(t)^k\,dt$
2	$\frac{1}{\sqrt{\pi}}$
3	$\frac{3}{2\sqrt{\pi}}$
4	$\frac{6}{\pi\sqrt{\pi}}\mathrm{Tan}^{-1}\sqrt{2}$
5	$\frac{10}{\sqrt{\pi}}\left(\frac{3}{2\pi}\mathrm{Tan}^{-1}\sqrt{2}-\frac{1}{4}\right)$

was determined independently, for different values of k, then Fig. 9 shows that n^* decreases as k increases. The highlighted points in Fig. 9 indicate the optimal sample size (n^*) for a given number of treatments (k). This behavior is to be expected since an increase in replicates eats into the deployment phase at a steeper rate when there are more treatments. Another finding is that the overall cumulative improvement is greater at the optimal point as k increases. Finally, it is also noteworthy that the overall experimental expenditure ($k \times n^*$) increases with an increase in k. We discuss this in greater detail following our derivations on n^*. While the optimal n^* can be inferred for a given environment, numerically (as done in Fig. 9), in this section, we analytically derive n^*. We identify n^* by solving for n in $\partial E(CI_{kn})/\partial n = 0$. Differentiating Eq. (19), we get

$$\frac{\sigma_e^2\sigma_m^2(T-kn)\left(\int_{-\infty}^{\infty}\frac{2^{1/2-k}ke^{-(t^2/2)}t\mathrm{erfc}\left(-\frac{t}{\sqrt{2}}\right)^{k-1}}{\sqrt{\pi}}\,dt\right)}{4n^2\left(\frac{\sigma_e^2}{n}+\frac{\sigma_m^2}{2}\right)^{3/2}} \qquad (20)$$

$$-\frac{k\sigma_m^2\left(\int_{-\infty}^{\infty}\frac{2^{1/2-k}ke^{-(t^2/2)}t\mathrm{erfc}\left(-\frac{t}{\sqrt{2}}\right)^{k-1}}{\sqrt{\pi}}\,dt\right)}{2\sqrt{\frac{\sigma_e^2}{n}+\frac{\sigma_m^2}{2}}}=0 \qquad (21)$$

Letting $\Omega = \sigma_m/\sigma_e$, we have

$$n^* = \frac{-3+\sqrt{9+\frac{4T}{k}\Omega^2}}{2\Omega^2} \qquad (22)$$

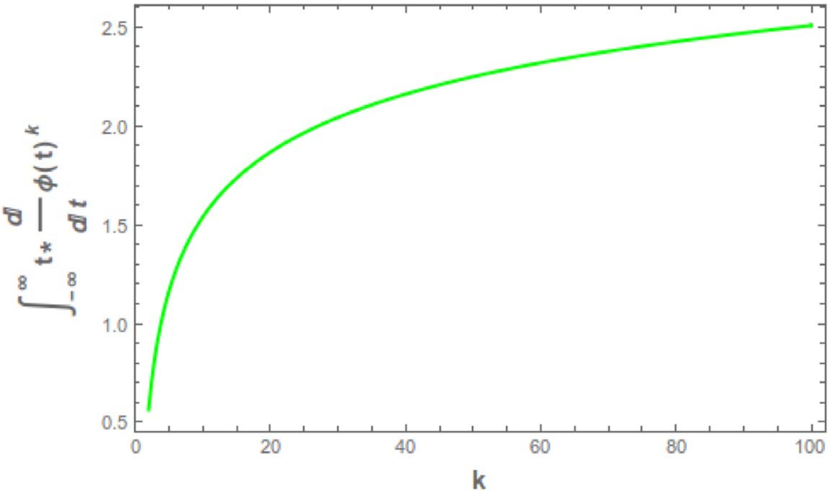

Fig. 8 Constant coefficient across different number of treatments

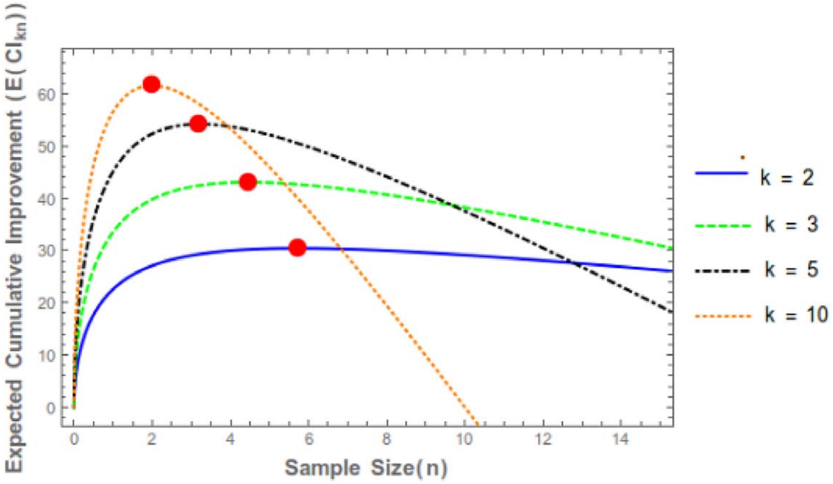

Fig. 9 Cumulative improvement across different sample size for $T = 100$ and $\Omega = 1$

It can be seen that Eq. (22) is not a function of $\int_{-\infty}^{\infty} t(d/dt)(\Phi(t)^k)\,dt$ (although cumulative improvement is influenced by this term). This is owing to the fact that this term is not a function of n and can be thought of as a constant for a given k. Also, as expected, Eq. (22) simplifies to Eq. (9) for $k=2$. This corroborates the claim that the optimal point for the minimization of cumulative regret and the maximization of cumulative improvement are the same. A closer look at Eq. (22) shows that the experimental budget $k \times n^*$ will always increase for an increase in k (despite the fact that n^* decreases with an increase in k). This is true for all values of T and Ω. It is also interesting to note that under no circumstances can the overall experimental budget ($n^* \times k$) as a percentage of the trial horizon (T) exceed $\frac{1}{3}$, and this happens when $k \rightarrow \infty$, or $\Omega \rightarrow 0$, or $T \rightarrow 0$ (It is important to note that there are real-world constraints such as $k \times n^*$ needing to be less than T or that n, k and T needing to be integers. Our formulations do not impose these practical consideration with the intent of understanding the true interrelationship between the variables and their edge conditions). In other words, $n^* k/T \rightarrow \frac{1}{3}$ as $k \rightarrow \infty$, or $\Omega \rightarrow 0$, or $T \rightarrow 0$. Similarly, $n^* k/T \rightarrow 0$ when $k \rightarrow 0$, or $\Omega \rightarrow \infty$, or $T \rightarrow \infty$. These edge conditions are important findings of the study.

5 Conclusions and Future Work

This paper recommends an analytical framework to make predetermined plans for the learning phase(s) of a project, where the system is live, and the scope of exploitation is defined. In essence, it helps practitioners make decisions on the allocation of time and budget for the experimental phase in this setting. The study achieves this through theoretical derivations of the optimal sample under different assumptions of the underlying environment. We demonstrate, in Sec. 4.3, that the degree of experimentation as a percentage of the trial horizon ranges from 0% to 33.33%, for any value of T, k, and Φ for when the priors are equal. These findings should equally apply to the base case in Sec. 3 and are also seen in setups explored in Sec. 4.1.

The estimate of $\Omega = \sigma_m / \sigma_e$ is an important input to determine the experimental budget. Any inaccuracy in the estimate would change the optimal budget and would, in turn, affect the performance results we discuss above. There are two major insights that our study offers in this regard. First, the potential loss due to moderate inaccuracies in this estimate is relatively minor. By contrast, differences in the implementation horizon, which is likely to be well-known, have a much greater impact on the experimental budget. For instance, when the true environment reflects an $\Omega = 1$, but the practitioner believes a different $\Omega = \frac{1}{2}$ or 2, this only causes the degree of experimentation to be off by approximately 6% on either side. In the worst case, the normalized expected cumulative regret ($NECR_{2n}$) the practitioner will experience is 13.6% as opposed to the 12% at optima. When this is compared to the baseline average $NECR_{2n}$ of 28%, experienced through a uniformly random experimental budget, the reduction in total improvement drops marginally from 73% to 68%. Whereas, as shown in Fig. 3, the sample size and performance significantly vary with T. The second insight, which follows from the results in Secs. 3 and 4, is that under-experimenting results in a larger loss than over-experimenting. In our example of using $\Omega = \frac{1}{2}$ or 2, instead of $\Omega = 1$, the pathological case occurs when the practitioner believes that $\Omega = 2$ and therefore under-experiments. This results in a $NECR_{2n} = 13.6\%$, whereas the regret experienced from over-experimenting due to an $\Omega = \frac{1}{2}$ is 12.9%.

Finally, quantifying the exact benefits from the proposed framework and comparing it with other alternative methods poses certain challenges, as the context is previously unexplored (to the authors' best knowledge). Hence, we do not have an established baseline approach for allocating a budget for one-shot experiments in the online framework, and it is unclear as to what other heuristics practitioners may employ. Despite this, a simple benchmark would be to compare the improvement in performance (lowering of expected cumulative regret) that can be availed by experimenting at the recommended optimal-level versus a uniformly randomly selected percentage of the horizon. For the two-treatment setting we explored in Sec. 3, we can see that when $\Omega = 1$, the lowest normalized expected cumulative regret ($NECR_{2n}$) is approximately 12% whereas a uniformly random experimental budget allocation from 0 to 100% would have yielded a regret of 28% on average. This is a reduction in regret of 16% in absolute terms, and in relative terms translates to a further improvement in performance by 73% (using $NECR_{2n} = 50\%$ as the worst case which is realized when the experimentation is 0% or 100%). Similarly, we see absolute reductions in $NECR_{2n}$ of 12% and 20% for $\Omega = \frac{1}{2}$ and 2, respectively. This translates to an improvement in the performance of 64% and 87%. Our extension discussed in Sec. 4.1 shows similar ranges of performance.

In terms of future work, a promising next step would be to explore a deployment that can happen over multiple phases. Our base case studies two phases of deployment, where the first phase gets committed to experimentation, and the next phase picks a winner. Our extension in Sec. 4.2 explores the idea of having unequal priors. If these unequal priors were to be modeled as the product of preceding phases of experimentation in an online environment, it would lend to a generalization where we could model more than two phases. Here, an algorithm would generate plans for a fixed number (m) of phases, over which the T units need to be deployed. Clearly, when $m = T$, our problem statement becomes the same as the multi-armed bandit setup. We currently explore the $m = 2$ setting. We believe that generating algorithms for settings where $2 \leq m \leq T$ would be a contribution of immense practical relevance. Another area of inquiry could be to explore solutions that make no distributional assumptions. It is also possible to consider an extension to continuous-value case for the treatments, where we could model this as an analogy to the infinite armed bandit setting. Finally, it would be of interest to study the various assumptions that a practitioner can make in a given environment and needs to make for a given method of analysis. For instance, in this work, we explore assumptions on the distributions of both the true means of the treatments, as well as the noise. One could study the various modes through which a practitioner can ascertain these values. It would also be important to understand if the knowledge of the uncertainty in the estimates can be used to make an experimental strategy more robust.

Acknowledgment

This research was partially funded by support from the Robert Bosch Center for Data Science and Artificial Intelligence (RBCDSAI), IIT Madras.

References

[1] Otto, K. N., and Antonsson, E. K., 1993, "Extensions to the Taguchi Method of Product Design," ASME J. Mech. Des., **115**(1), pp. 5–13.

[2] Siang, J. K. K., Chia, P. Z., Koronis, G., and Silva, A., 2018, "Exploring the Use of a Full Factorial Design of Experiment to Study Design Briefs for Creative Ideation," ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, New York, p. V007T06A008.

[3] Frey, D. D., and Jugulum, R., 2003, "How One-Factor-at-a-Time Experimentation Can Lead to Greater Improvements Than Orthogonal Arrays," ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, New York, pp. 505–513.

[4] Sasena, M., Parkinson, M., Goovaerts, P., Papalambros, P., and Reed, M., 2002, "Adaptive Experimental Design Applied to an Ergonomics Testing Procedure DAC-34091," Proceedings of the 2002 ASME DETC Conferences, Montreal, Canada, Sept. 29–Oct. 2.

[5] Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N.-H., 2010, "Adaptive Designs of Experiments for Accurate Approximation of a Target Region," ASME J. Mech. Des., **132**(7), p. 071008.

[6] Sutton, R. S., and Barto, A. G., 1998, *Reinforcement Learning: An Introduction*, Vol. 1, MIT Press, Cambridge.

[7] Burtini, G., Loeppky, J., and Lawrence, R., 2015, "A Survey of Online Experiment Design With the Stochastic Multi-Armed Bandit," arXiv:1510.00757 .

[8] Montgomery, D. C., 2017, *Design and Analysis of Experiments*, John Wiley & Sons, Hoboken, NJ.

[9] Adcock, C. J., 1997, “Sample Size Determination: A Review,” J. R. Stat. Soc.: Ser. D (Statistician), **46**(2), pp. 261–283.

[10] Lenth, R. V., 2001, “Some Practical Guidelines for Effective Sample Size Determination,” Am. Statistician, **55**(3), pp. 187–193.

[11] Schlaifer, R., and Raiffa, H., 1961, *Applied Statistical Decision Theory*, Division of Research, Graduate School of Business Administration, Harvard.

[12] Fraser, D. A. S., and Guttman, I., 1956, “Tolerance Regions,” Ann. Math. Stat., **27**(1), pp. 162–179.

[13] Willan, A. R., 2008, “Optimal Sample Size Determinations From An Industry Perspective Based on the Expected Value of Information,” Clin. Trials, **5**(6), pp. 587–594.

[14] Chaloner, K., and Verdinelli, I., 1995, “Bayesian Experimental Design: A Review,” Stat. Sci., **10**(3), pp. 273–304.

[15] Even-Dar, E., Mannor, S., and Mansour, Y., 2006, “Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems,” J. Mach. Learn. Res., **7**(Jun), pp. 1079–1105.

[16] Ciarleglio, M. M., Arendt, C. D., and Peduzzi, P. N., 2016, “Selection of the Effect Size for Sample Size Determination for a Continuous Response in a Superiority Clinical Trial Using a Hybrid Classical and Bayesian Procedure,” Clin. Trials, **13**(3), pp. 275–285.

[17] Schönbrodt, F. D., and Wagenmakers, E. -J., 2018, “Bayes Factor Design Analysis: Planning for Compelling Evidence,” Psychon. Bull. Rev., **25**(1), pp. 128–142.

[18] Sudarsanam, N., Pitchai Kannu, B., and Frey, D. D., 2019, “Optimal Replicates for Designed Experiments Under the Online Framework,” Res. Eng. Des., **30**(3), pp. 363–379.

[19] Bellman, R., 1956, “Dynamic Programming and Lagrange Multipliers,” Proc. Natl. Acad. Sci., **42**(10), pp. 767–769.

[20] Robbins, H., 1985, *Herbert Robbins Selected Papers*, T. L. Lai and D. Siegmund, eds., Springer, New York, pp. 169–177.

[21] Thompson, W. R., 1933, “On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples,” Biometrika, **25**(3/4), pp. 285–294.

[22] Auer, P., Cesa-Bianchi, N., and Fischer, P., 2002, “Finite-Time Analysis of the Multiarmed Bandit Problem,” Mach. Learn., **47**(2–3), pp. 235–256.

[23] Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., Grünwald, P., Hazan, E., and Kale, S., 2015, “Batched Bandit Problems,” Proceedings of the 28th Conference on Learning Theory, Paris, July 3–6, p. 1456.

[24] Auer, P., and Ortner, R., 2010, “UCB Revisited: Improved Regret Bounds for the Stochastic Multi-Armed Bandit Problem,” Period. Math. Hungarica, **61**(1–2), pp. 55–65.

[25] Pham-Gia, T., and Turkkan, N., 1992, “Sample Size Determination in Bayesian Analysis,” Statistician, **41**(4), pp. 389–397.

[26] Joseph, L., and Belisle, P., 1997, “Bayesian Sample Size Determination for Normal Means and Differences Between Normal Means,” J. R. Stat. Soc.: Ser. D (Statistician), **46**(2), pp. 209–226.

[27] Machin, D., Campbell, M. J., Tan, S.-B., and Tan, S.-H., 2011, *Sample Size Tables for Clinical Studies*, John Wiley & Sons, Englewood, NJ.

[28] Stein, C., 1945, “A Two-Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance,” Ann. Math. Stat., **16**(3), pp. 243–258.

[29] Simon, R., Wittes, R., and Ellenberg, S., 1985, “Randomized Phase II Clinical Trials,” Cancer Treat. Rep., **69**(12), pp. 1375–1381.

[30] Simon, R., 1989, “Optimal Two-Stage Designs for Phase II Clinical Trials,” Contemp. Clin. Trials, **10**(1), pp. 1–10.

[31] Frey, D. D., and Wang, H., 2006, “Adaptive One-Dactor-at-a-Time Experimentation and Expected Value of Improvement,” Technometrics, **48**(3), pp. 418–431.

[32] Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N., 1992, *A First Course in Order Statistics*, Vol. 54, SIAM, Philadelphia, PA.