Operations Research

Simple Bayesian Algorithms for Best-Arm Identification

Daniel Russo

OPERATIONS RESEARCH
Vol. 68, No. 6, November–December 2020, pp. 1625–1647
ISSN 0030-364X (print), ISSN 1526-5463 (online)
http://pubsonline.informs.org/journal/opre

Methods

Simple Bayesian Algorithms for Best-Arm Identification

Daniel Russo[a]

[a] Columbia University, New York, New York 10027

Contact: djr2174@gsb.columbia.edu, http://orcid.org/0000-0001-5926-8624 (DR)

Received: December 12, 2016
Revised: May 4, 2018; November 23, 2018; April 2, 2019
Accepted: April 10, 2019
Published Online in Articles in Advance: April 16, 2020

Subject Classifications: decision analysis: sequential; statistics: Bayesian
Area of Review: Methods

https://doi.org/10.1287/opre.2019.1911

Abstract. This paper considers the optimal adaptive allocation of measurement effort for identifying the best among a finite set of options or designs. An experimenter sequentially chooses designs to measure and observes noisy signals of their quality with the goal of confidently identifying the best design after a small number of measurements. This paper proposes three simple and intuitive Bayesian algorithms for adaptively allocating measurement effort and formalizes a sense in which these seemingly naive rules are the best possible. One proposal is top-two probability sampling, which computes the two designs with the highest posterior probability of being optimal and then randomizes to select among these two. One is a variant of top-two sampling that considers not only the probability that a design is optimal, but the expected amount by which its quality exceeds that of other designs. The final algorithm is a modified version of Thompson sampling that is tailored for identifying the best design. We prove that these simple algorithms satisfy a sharp optimality property. In a frequentist setting where the true quality of the designs is fixed, one hopes that the posterior definitively identifies the optimal design, in the sense that that the posterior probability assigned to the event that some other design is optimal converges to zero as measurements are collected. We show that under the proposed algorithms, this convergence occurs at an exponential rate, and the corresponding exponent is the best possible among all allocation rules. It should be highlighted that the proposed algorithms depend on a single tuning parameter, which determines the probability used when randomizing among the top-two designs. Attaining the optimal rate of posterior convergence requires either that this parameter is set optimally or is tuned adaptively toward the optimal value. The paper goes further, characterizing the exponent attained on any problem instance and for any value of the tunable parameter. This exponent is interpreted as being optimal among a constrained class of allocation rules. Finally, considerable robustness to this parameter is established through numerical experiments and theoretical results. When this parameter is set to 1/2, the exponent attained is within a factor of 2 of best possible across all problem instances.

History: First Place Winner, Junior Faculty Forum Paper Competition, 2016.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2019.1911.

Keywords: multiarmed bandit • ranking and selection • Bayesian

1. Introduction

This paper considers the optimal adaptive allocation of measurement effort in order to identify the best among a finite set of options or designs. An experimenter sequentially chooses designs to measure and observes independent noisy signals of their quality. The goal is to allocate measurement effort intelligently so that the best design can be identified confidently after a small number of measurements. Just as the multiarmed bandit problem crystallizes the tradeoff between exploration and exploitation in sequential decision making, this "pure-exploration" problem crystallizes the challenge of efficiently gathering information before committing to a final decision. It serves as a fundamental abstraction of issues faced in many practical settings. For example:

- *Efficient A/B/C Testing*: An e-commerce platform is considering a change to its website and would like to identify the best-performing candidate among many potential new designs. To do this, the platform runs an experiment, displaying different designs to different users who visit the site. How should the platform decide what percentage of traffic to allocate to each website design?
- *Simulation Optimization*: An engineer would like to identify the best-performing aircraft design among several proposals. She has access to a realistic simulator through which she can assess the quality of the designs, but each simulation trial is very time consuming

and produces only noisy output. How should she allocate simulation effort among the designs?

- *Design of Clinical Trials*: A medical research organization would like to find the most effective treatment out of several promising candidates. They run a clinical trial in which they experiment with the treatments. The results of the study may influence practice for many years to come, so it is worth reaching a definitive conclusion. At the same time, clinical trials are extremely expensive, and careful experimentation can help to mitigate the associated costs.[1] Multiarmed bandit models of clinical trials date back to Thompson (1933), but bandit algorithms lack statistical power in detecting the best treatment at the end of the trial (Villar et al. 2015). Can we develop adaptive rules with better performance?

We study Bayesian algorithms for adaptively allocating measurement effort. Each begins with a prior distribution over the unknown quality of the designs. The experimenter learns as measurements are gathered, and beliefs are updated to form a posterior distribution. This posterior distribution gives a principled mechanism for reasoning about the uncertain quality of designs and for assessing the probability that any given design is optimal. By formulating this problem as a Markov decision process whose state-space tracks posterior beliefs about the true quality of each design, dynamic programming could, in principle, be used to optimize many natural measures of performance. Unfortunately, computing or even storing an optimal policy is usually infeasible due to the curse of dimensionality. Instead, this work *proposes three simple and intuitive rules for adaptively allocating measurement effort and, by characterizing fundamental limits on the performance of any algorithm, formalizes a sense in which these seemingly naïve rules are the best possible.*

The first algorithm we propose is called *top-two probability sampling*. It computes at each time step the two designs with the highest posterior probability of being optimal. It then randomly chooses among them, selecting the design that appears most likely to be optimal with some fixed probability, and selecting the second most likely otherwise. Beliefs are updated as observations are collected, so the top-two designs change over time. The long-run fraction of measurement effort allocated to each design depends on the true quality of the designs and the distribution of observation noise. *Top-two value sampling* proceeds in a similar manner, but in selecting the top-two designs, it considers not only the probability that a design is optimal, but the expected amount by which its quality exceeds that of other designs. The final algorithm we propose is a top-two sampling version of the *Thompson sampling* algorithm for multiarmed bandits. Thompson sampling has attracted a great deal of recent interest in both academia and industry (Graepel et al. 2010, Chapelle and Li 2011, Agrawal and Goyal 2012, Kaufmann et al. 2012, Tang et al. 2013, Gopalan et al. 2014, Scott 2016, Russo and Van Roy 2017), but it is designed to maximize the cumulative reward earned while sampling. As a result, in the long run, it allocates almost all effort to measuring the estimated-best design and requires a huge number of total measurements to certify that none of the alternative designs offer better performance. We introduce a natural top-two variant of Thompson sampling that avoids this issue and, as a result, offers vastly superior performance for the best-arm identification problem.

Remarkably, these simple heuristic algorithms satisfy a strong optimality property. Our analysis focuses on frequentist consistency and rate convergence of the posterior distribution (see, e.g., Freedman 1963) and therefore takes place in a setting where the true quality of the designs is fixed, but unknown to the experimenter. One hopes that as measurements are collected, the posterior distribution definitively identifies the true best design, in the sense that the posterior probability assigned to the event that some other design is optimal converges to zero. We show that under the proposed algorithms, this convergence occurs at an exponential rate, characterize the exponent attained for each problem instance, and relate this to the best possible exponent among allocation rules.

To make a precise statement, it is important to highlight that the top-two algorithms described above depend on a tunable parameter; each method identifies the top-two designs and then flips a biased coin to decide which of these to sample. The paper's theoretical results offer a fairly complete characterization of the asymptotic performance of these algorithms and are summarized more precisely below.

(1) **Optimality with tuning:** For any problem instance and any choice of tuning parameter, the proposed top-two algorithms attain an exponential rate of posterior convergence. This exponent is carefully characterized. If the tuning parameter is set optimally, the exponent is optimal among all possible adaptive allocation rules. Moreover, it is possible to attain this rate of convergence by adaptively adjusting the tuning parameter.

(2) **Robustness with an unbiased coin:** Uniformly across problem instances, the exponent attained by top-two sampling with an unbiased coin is within a factor of two of what could be attained by an optimal allocation rule. This robustness is further validated through numerical experiments: Across 14 problem instances, top-two Thompson sampling with an unbiased coin offers similar performance to a version of top-two Thompson sampling that is applied with the best tuning parameter for that particular problem setting.

(3) **Optimality among a restricted class of allocation rules for any tuning parameter:** To simplify the discussion, imagine that top-two sampling is

applied with an unbiased coin. Then, as the number of measurements tends to infinity, exactly half of the measurement effort is allocated to the best design. Now, consider any possible adaptive allocation rule, which, like top-two sampling, allocates half of the measurement effort to the true best design asymptotically. There is no problem instance for which this alternative algorithm attains an exponential rate of posterior convergence exceeding that of the proposed top-two sampling algorithms. An analogous result applies when a biased coin is used.

It is worth elaborating on the third result described above, as it is the main insight that prompted this paper. We face the problem of adaptively allocating measurements among k competing designs. We can imagine decomposing this problem into two parts: First, the experimenter chooses which fraction of measurements to dedicate to what is believed to be the best design, and, second, given this choice, she chooses how to adaptively allocate remaining measurements among the $k-1$ competing designs. Roughly speaking, this paper shows that the allocation among the remaining $k-1$ designs is handled automatically and optimally by very simple top-two sampling algorithms. This offers substantial new insight into the structure of best-arm identification problems and effectively reduces the problem to the choice of a single tuning parameter—the bias of the coin used by the top-two sampling algorithms. The paper establishes a surprising degree of robustness to this tuning parameter and shows that it is possible to attain a fully optimal exponent by setting it adaptively. However, the proposed tuning method is complex, spoiling some of the elegance of the top-two sampling algorithms. The search for simpler methods stands as an interesting open question.

Finally, it should be highlighted that the performance metric studied in this paper—concerning the frequentist rate of convergence of the posterior—is *not* the same as the widely studied probability of incorrect selection metric. The claims in this paper do not imply the probability of incorrect selection converges at an optimal rate, and, in fact, the current literature does not provide such a result for any adaptive algorithm. Extending the theory in this paper to cover more conventional performance metrics is an important direction for future work. Thankfully, the paper's analysis does seem to have deep connections with other performance metrics. The interested reader can find a more careful discussion in Online Appendix EC.2 or in a follow-up to this paper by Qin et al. (2017). That paper provides similar asymptotically optimal guarantees for a top-two sampling algorithm applied to best-arm identification in the so-called "fixed-confidence" setting, which is a purely frequentist problem formulation that is widely studied in the literature.

1.1. Main Contributions

This paper makes both algorithmic and theoretical contributions. On the algorithmic side, we develop three new adaptive measurement rules. The top-two Thompson sampling rule, in particular, could have an immediate impact in application areas where Thompson sampling is already in use. For example, there are various reports of Thompson sampling being used in A/B testing (Scott 2016) and in clinical trials (Berry 2004). But practitioners in these domains typically hope to commit to a decision after a definitive period of experimentation, and top-two Thompson sampling can greatly reduce the number of measurements required to do so. In addition, because of their simplicity, the proposed allocation rules can be easily adapted to treat problems beyond the scope of this paper's problem formulation. See Section 8 for examples.

The paper also makes several theoretical contributions. Most importantly, it is of broad scientific interest to understand when very simple measurement strategies are the best possible. This paper provides sharp links between these top-two sampling rules and the limits of performance under any adaptive algorithm. In establishing these results, we exactly characterize the optimal rate of posterior convergence attainable by an adaptive algorithm and provide interpretable bounds on this rate when measurement distributions are sub-Gaussian. The analysis also provides several intermediate results, which may be of independent interest, including establishing consistency and exponential rates of convergence for posterior distributions with nonconjugate priors and under adaptive measurement rules. It should be highlighted, however, that the results do require some strong regularity properties on the prior distribution and, in particular, only apply to priors defined over a compact set.

1.2. Related Literature

Sequential Bayesian Best-Arm Identification. There is a sophisticated literature on algorithms for Bayesian multiarmed bandit problems. In discounted bandit problems with independent arms, Gittins indices characterize the Bayes optimal policy (Gittins and Jones 1974, Gittins 1979). Moreover, a variety of simpler Bayesian allocation rules have been developed, including Bayesian upper-confidence bound algorithms (Kaufmann et al. 2012, Srinivas et al. 2012, Kaufmann 2018), Thompson sampling (Agrawal and Goyal 2012, Korda et al. 2013, Gopalan et al. 2014, Ferreira et al. 2018), information-directed sampling (Russo and Van Roy 2017), the knowledge gradient (Ryzhov et al. 2012), and optimistic Gittins indices (Gutin and Farias 2016). These heuristic algorithms can be applied effectively to complicated learning problems beyond the specialized settings in which the Gittins index theorem holds, have been shown to have strong performance in simulation, and have theoretical performance guarantees. In several cases, they are known to attain sharp asymptotic limits on the performance of any adaptive algorithm due to Lai and Robbins (1985).

The pure-exploration problem studied in this paper is not nearly as well understood. Recent work has cast this problem in a decision-theoretic framework (Chick and Gans 2009). However, because the conditions required for the Gittins index theorem do not hold, computing an optimal policy via dynamic programming is generally infeasible due to the curse of dimensionality. Papers in this area typically focus on problems with Gaussian observations and priors. They formulate simpler problems that can be solved exactly—like a problem where only a single measurement can be gathered (Gupta and Miescke 1996, Frazier et al. 2008, Chick et al. 2010) or a continuous-time problem with only two alternatives (Chick and Frazier 2012)—and then extend those solutions heuristically to build measurement and stopping rules in more general settings.

For problems with Gaussian priors and noise distributions, the expected-improvement (EI) algorithm is a popular Bayesian approach to sequential information gathering. Interesting recent work by Ryzhov (2016) studies the long-run distribution of measurement effort allocated by the expected improvement and shows that this is related to the optimal-computing-budget allocation of Chen et al. (2000). This contribution is very similar in spirit to this paper, as it relates the long-run behavior of a simple Bayesian measurement strategy to a notion of an approximately optimal allocation. Unfortunately, EI cannot match the performance guarantees in this paper. In fact, under EI, the posterior converges only at a polynomial rate, instead of the exponential rate attained by the algorithms proposed here and by the optimal-computing-budget allocation (OCBA). See Online Appendix EC.4 for a more precise discussion.

Classical Ranking and Selection. The problem of identifying the best system has been studied for many decades under the names *ranking and selection* or *ordinal optimization*. A full review of this literature is beyond the scope of this article. See Kim and Nelson (2006), Kim and Nelson (2007), or Hong et al. (2015) for thorough reviews. Part of this literature focuses on a problem called subset selection, where the goal is not to identify the best design, but to find a fairly small subset of designs that is guaranteed to contain the best design. Beginning with Bechhofer and Sobel (1954), many papers have focused on an indifference zone formulation, where, for user-specified $\epsilon, \delta > 0$, the goal is to guarantee with probability at least $1-\delta$ the algorithm returns the true arm mean, as long as no suboptimal arm is within ϵ of optimal. Assuming that measurement noise is Gaussian with known variance σ^2, one can guarantee this indifference-zone criterion by gathering $O\big((\sigma k/\epsilon^2)\log(k/\delta)\big)$ total measurements, divided equally among the k designs, and then returning the design with the highest empirical mean. For the case of unknown variances, Rinott (1978) proposes a two-stage procedure, where the first stage is used to estimate the variance of each population, and the number of samples collected from each design in the second stage is scaled by its estimated standard deviation. In the machine learning literature, Even-Dar et al. (2002) studies the number of samples required by algorithms delivering ϵ–PAC guarantees. Such algorithms are sometimes said to ensure a specified *probability of good selection* in the terminology of the simulation-optimization literature, a strictly stronger guarantee than an indifference-zone guarantee (Ni et al. 2017). Even-Dar et al. (2002) show that when measurement noise is uniformly bounded, an ϵ–PAC guarantee is satisfied by a sequential elimination strategy that uses only $O\big((k/\epsilon^2)\log(1/\delta)\big)$ samples on average. Mannor and Tsitsiklis (2004) provide a matching lower bound. Similar to minimax bounds, this shows that the upper bound of Even-Dar et al. (2002) is tight, up to a constant factor, for a certain worst-case problem instance. Indifference-zone formulations of ranking and selection problems remains an area of active research. See, for example, Fan et al. (2016) and some of the references therein.

Since Paulson (1964), many authors have sought to reduce the number of samples required on easier problem instances by designing algorithms that sequentially eliminate arms once they are determined to be suboptimal with high confidence. See the recent work of Frazier (2014) and the references therein. However, in a sense described below, Jennison et al. (1982) show formally that there are problems with Gaussian observations where any sequential-elimination algorithm will require substantially more samples than optimal adaptive allocation rules. See Section 8 for modified top-two sampling algorithms designed for an indifference zone criterion.

The Asymptotic Complexity of Best-Arm Identification. We described attainable rates of performance on a worst-case problem instance characterized by Even-Dar et al. (2002) and Mannor and Tsitsiklis (2004). A great deal of work has sought “problem-dependent” bounds, which reveal that the best arm can be identified more rapidly when the true problem instance is easier. This is the case, for example, when some arms are of very low quality and can be distinguished from the best by using a small number of measurements.

Asymptotic measures of the complexity of best-arm identification appear to have been derived independently in statistics (Chernoff 1959, Jennison et al. 1982), simulation optimization (Glynn and Juneja 2004), and, concurrently with this paper, in the machine learning literature (Garivier and Kaufmann 2016). Each of these papers studies a slightly different objective, but each captures a notion of the number of samples required to identify the best arm as a function of the problem instance—that is, as a function the number of designs, each design's true quality, and the distribution of measurement noise.

Glynn and Juneja (2004) build on the OCBA of Chen et al. (2000) to provide a rigorous large-deviations derivation of the optimal fixed allocation. In particular, assuming the design with the highest empirical mean is returned, there is a fixed allocation under which the probability of incorrect selection decays exponentially, and the exponent is optimal under all fixed-allocation rules. The setting studied by this paper is often called the "fixed-budget" setting in the recent multiarmed bandit literature. Unfortunately, it may be difficult to implement the allocation in Glynn and Juneja (2004) without additional prior knowledge, and so it is unclear whether their large deviations exponent is attainable by an adaptive algorithm. Later work by Glynn and Juneja (2015) provides a discussion of this issue.

This paper was highly influenced by a classic paper by Chernoff (1959) on the sequential design of experiments for binary hypothesis testing. Chernoff's asymptotic derivations give great insight into best-arm identification, which can be formulated as a multiple-hypothesis testing problem with sequentially chosen experiments, but, surprisingly, this connection does not seem to be discussed in the literature. Chernoff looks at a different scaling than Glynn and Juneja (2004). Instead of taking the budget of available measurements to infinity, he allows the algorithm to stop and declare the hypothesis true or false at any time, but takes the cost of gathering measurements to zero, while the cost of an incorrect terminal decision stays fixed. He constructs rules that minimize expected total costs in this limit. Chernoff makes restrictive technical assumptions, some of which have been removed in subsequent work (Albert 1961, Kiefer and Sacks 1963, Keener 1984, Naghshvar et al. 2013, Nitinawarat et al. 2013).

Jennison et al. (1982) study an indifference-zone formulation of the problem of identifying the best design. Like Chernoff (1959), they allow the algorithm to stop and return an estimate of the best arm at any time, but rather than penalize incorrect decisions, they require that the probability correct selection (PCS) exceeds $1 - \delta > 0$ for every problem instance. Intuitively, the expected number of samples required by an algorithm satisfying this PCS constraint must tend to infinity as $\delta \to 0$. In the case of Gaussian measurement noise, Jennison et al. (1982) characterize the optimal asymptotic scaling of expected number of samples in this limit. The recent multiarmed bandit literature refers to this formulation as the fixed-confidence setting.

A large body of work in the recent machine learning literature has sought to characterize various notions of the complexity of best-arm identification (Even-Dar et al. 2002, Mannor and Tsitsiklis 2004, Audibert et al. 2010, Gabillon et al. 2012, Karnin et al. 2013, Jamieson and Nowak 2014). However, upper and lower bounds match up only to constant or logarithmic factors, and only for particular hard problem instances. Substantial progress was presented by Kaufmann and Kalyanakrishnan (2013) and Kaufmann et al. (2014), who seek to exactly characterize the asymptotic complexity of identifying the best arm in both the fixed-budget and fixed-confidence settings. Still, the upper and lower bounds presented there do not match. A short abstract of the current paper appeared in the 2016 Conference on Learning Theory. In the same conference, independent work by Garivier and Kaufmann (2016) provided matching upper and lower bounds on the complexity of identifying the best arm in the fixed-confidence setting. Like the present paper, but unlike Jennison et al. (1982), these results apply whenever observation distributions are in the exponential family and do not require an indifference zone.

The current paper looks at a different measure. We study a frequentist setting in which the true quality of each design is fixed and characterize the rate of posterior convergence attainable for each problem instance. We also describe, as a function of the problem instance, the long-run fraction of measurement effort allocated to each design by any algorithm attaining this rate of convergence. These asymptotic limits turn out to be closely related to some of the aforementioned results. In particular, the optimal exponent given in Subsection 6.4 mirrors the complexity measure of Chernoff (1959). In the same subsection, this exponent is then simplified into a form derived for Gaussian distributions by Jennison et al. (1982).

Optimal Budget Allocations. Although the complexity measure we derive is similar to past work, the proposed algorithms differ substantially. The allocation rules proposed by Chernoff (1959), Jennison et al. (1982), and Glynn and Juneja (2004) are essentially developed as a means of proving that certain rates are attainable asymptotically. To derive these policies, the authors begin with a thought experiment: Assuming the experimenter actually knew the true quality of every arm, what proportion of measurements should she allocate to each arm in order to gather the most definitive evidence concerning the identity of the optimal arm? One approach to constructing such rules in practice is to use some fraction of samples to estimate the arm means and then apply the asymptotically optimal sampling proportions assuming these estimates to be correct. Such an approach dates back to at least the work of Kiefer and Sacks (1963), which followed Chernoff's work on the sequential design of experiments.

Early authors made a point to highlight limitations of such an approach. Jennison et al. (1982) writes that their proposed procedures "typically…do not have good small sample size properties. A better procedure would have several stages and a more sophisticated sampling rule." In a 1975 review of the sequential design of experiments, Chernoff (1975) notes that asymptotic approaches to the optimal sequential design of experiments had been fairly successful in circumventing the need to compute Bayesian optimal designs via dynamic programming, but "the approach is very coarse for moderate sample size problems." He writes that two-stage procedures of Kiefer and Sacks (1963), "sidestep the issue of how to experiment in the early stages," while constructing the optimal allocations based on point estimates "treats estimates of $\boldsymbol{\theta}$ based on a few observations with as much respect as that based on many observations."

Closely related to these approaches is a large body of work on optimal computing budget allocations (Chen et al. 2000). Most of this literature studies problems with Gaussian observations. They derive an approximation to the optimal sampling proportions as they are presented in Glynn and Juneja (2004), which appears to simplify computation. This allocation is often stated to be optimal as the number of arms grows large; more rigorous results to this effect are established in interesting work by Pasupathy et al. (2015), who shows that the sampling ratios of the OCBA coincide with those of Glynn and Juneja (2004) in the limit of a sequence of problem instances in which the number of arms tends to infinity but all suboptimal arms' means are bounded away from optimal by a fixed constant. Optimal budget allocations have been extended in various directions—for example, to address Bayesian expected loss objectives (Chick and Inoue 2001), the problem of identifying an optimum subject to stochastic constraints (Hunter and Pasupathy 2013), and the problem of identifying the top m alternatives (Chen et al. 2008). See Chen et al. (2015) for a more thorough review.

In this paper, we study simple adaptive allocation rules, which, ostensibly, have no relation to the asymptotic calculations used to derive these optimal budget allocations. The main insight is that these simple algorithms automatically adapt their measurement effort in such a way that their long-run behavior is deeply linked to the ratios suggested in the work of Chernoff (1959) and Jennison et al. (1982). A major advantage of top-two sampling algorithms, however, is that asymptotic analysis is used only to give insight into the algorithms, and any approximations have no impact on their practical performance. A suite of experiments in Section 7 suggest that the approach can substantially outperform the optimal allocations derived from asymptotic theory.

2. Problem Formulation

Consider the problem of efficiently identifying the best among a finite set of designs based on noisy sequential measurements of their quality. At each time $n \in \mathbb{N}$, a decision maker chooses to measure the design $I_n \in \{1, \ldots, k\}$ and observes a measurement Y_{n,I_n}. The measurement $Y_{n,i} \in \mathbb{R}$ associated with design i and time n is drawn from a fixed, unknown probability distribution, and the vector $\boldsymbol{Y}_n \triangleq (Y_{n,1}, \ldots, Y_{n,k})$ is drawn independently across time. The decision maker chooses a *policy*, or *adaptive allocation rule*, which is a (possibly randomized) rule for choosing a design I_n to measure as a function of past observations $I_1, Y_{1,I_1}, \ldots I_{n-1}, Y_{n-1,I_{n-1}}$. The goal is to efficiently identify the design with highest mean.

We will restrict attention to problems where measurement distributions are in the canonical one-dimensional exponential family. The marginal distribution of the outcome $Y_{n,i}$ has density $p(y|\theta_i^*)$ with respect to a base measure ν, where $\theta_i^* \in \mathbb{R}$ is an unknown parameter associated with design i. This density takes the form

$$p(y|\theta) = b(y)\exp\{\theta T(y) - A(\theta)\} \qquad \theta \in \mathbb{R}, \qquad (1)$$

where b, T, and A are known functions, and $A(\theta)$ is assumed to be twice differentiable. We will assume that T is a strictly increasing function so that $\mu(\theta) \triangleq \int yp(y|\theta)dv(y)$ is a strictly increasing function of θ. Many common distributions can be written in this form, including Bernoulli, normal (with known variance), Poisson, exponential, chi-squared, and Pareto (with known minimal value).

Throughout the paper, $\boldsymbol{\theta}^* \triangleq (\theta_1^*, \ldots, \theta_k^*)$ will denote the unknown true parameter vector, and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ will be used to denote possible alternative parameter vectors. Let $I^* = \arg\max_{1 \le i \le k} \theta_i^*$ denote the unknown best design. We will assume throughout that $\theta_i^* \neq \theta_j^*$ for $i \neq j$ so that I^* is unique, although this can be relaxed by considering an indifference zone formulation where the goal is to identify an ϵ–optimal design, for some specified tolerance level $\epsilon > 0$.

2.1. Prior and Posterior Distributions

The policies studied in this paper make use of a prior distribution Π_1 over a set of possible parameters Θ

that contains $\boldsymbol{\theta}^*$. Based on a sequence of observations $(I_1, Y_{1,I_1}, \ldots, I_{n-1}, Y_{n-1,I_{n-1}})$, beliefs are updated to attain a posterior distribution Π_n. We assume Π_1 has density π_1 with respect to Lebesgue measure. In this case, the posterior distribution Π_n has corresponding density

$$\pi_n(\boldsymbol{\theta}) = \frac{\pi_1(\boldsymbol{\theta})L_{n-1}(\boldsymbol{\theta})}{\int_\Theta \pi_1(\boldsymbol{\theta}')L_{n-1}(\boldsymbol{\theta}')d\boldsymbol{\theta}'} \quad n \geq 2, \tag{2}$$

where

$$L_{n-1}(\boldsymbol{\theta}) = \prod_{l=1}^{n-1} p\big(Y_{l,I_l}|\theta_{I_l}\big),$$

is the likelihood function. Although this formulation enforces some technical restrictions to facilitate theoretical analysis, it allows for very general prior distributions and, in particular, allows for the quality of different designs to be correlated under the priors.

2.2. Optimal Action Probabilities

Let

$$\Theta_i \triangleq \left\{\boldsymbol{\theta} \in \Theta \middle| \theta_i > \max_{j \neq i} \theta_j\right\}$$

denote the set of parameters under which design i is optimal, and let

$$\alpha_{n,i} \triangleq \Pi_n(\Theta_i) = \int_{\Theta_i} \pi_n(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{3}$$

denote the posterior probability assigned to the event that action i is optimal. Our analysis will focus on $\Pi_n(\Theta^c_{I^*}) = 1 - \alpha_{I^*}$, which is the posterior probability assigned to the event that an action other than I^* is optimal. The next section will introduce policies under which $\Pi_n(\Theta^c_{I^*}) \to 0$ as $n \to \infty$, and the rate of convergence is essentially optimal. This kind of analysis, which looks at the long-run dynamics of the posterior in a frequentist model, has a rich history in statistics (see, e.g., Freedman 1963, Diaconis and Freedman 1986, Barron et al. 1999, Ghosal et al. 2000).

Let me again highlight that *this performance metric differs from the probability of incorrect selection*. To study the probability of incorrect selection, we would fix a decision rule, with a conventional choice being to return the arm $\hat{i}_n$ that generated the largest empirical mean-reward prior to time n. We would then study the rate at which $\mathbb{P}(\hat{i}_n \neq I^*)$ decays as n grows under the proposed procedure for sampling arms. This setup is called the fixed-budget setting in the multiarmed bandit literature. The techniques in this paper do not yield exponentially decaying bounds on this metric and cannot be easily extended to do so. However, there are deep connections with best-arm identification in the so-called fixed-confidence setting, which are described in Online Appendix EC.2 or Qin et al. (2017).

2.3. Further Notation

Before proceeding, we introduce some further notation. Let $\mathcal{F}_n$ denote the sigma algebra generated by $(I_1, Y_{1,I_1}, \ldots I_n, Y_{n,I_n})$. For all $i \in \{1, \ldots, k\}$ and $n \in \mathbb{N}$, define

$$\psi_{n,i} \triangleq \mathbb{P}(I_n = i|\mathcal{F}_{n-1}) \quad \Psi_{n,i} \triangleq \sum_{\ell=1}^{n} \psi_{\ell,i} \quad \overline{\psi}_{n,i} \triangleq n^{-1}\Psi_{n,i}.$$

Each of these measures the effort allocated to design i up to time n.

3. Algorithms

This section proposes three algorithms for allocating measurement effort. Each depends on a tuning parameter $\beta > 0$, which will sometimes be set to a default value of 1/2. Each algorithm is based on the same high-level principle. At every time step, each algorithm computes an estimate $\hat{I} \in \{1, \ldots, k\}$ of the optimal design and measures that with probability β. Otherwise, we consider a counterfactual: In the (possibly unlikely) event that $\hat{I}$ is not the best design, which alternative $\hat{J} \neq \hat{I}$ is most likely to be the best design? With probability $1 - \beta$, the algorithm measures the alternative $\hat{J}$. The algorithms differ in how they compute $\hat{I}$ and $\hat{J}$. The most computationally efficient is the modified version of Thompson sampling, under which $\hat{I}$ and $\hat{J}$ are themselves randomly sampled from a probability distribution.

We will see that asymptotically, all three algorithms allocate fraction β of measurement effort to measuring the estimated-best design, and the remaining fraction to gathering evidence about alternatives. The algorithms adjust how measurement effort is divided among these alternative designs as evidence is gathered, allocating less effort to measuring clearly inferior designs and greater effort to measuring designs that are more difficult to distinguish from the best.

3.1. Top-Two Probability Sampling

With probability β, the top-two probability sampling (TTPS) policy plays the action $\hat{I}_n = \arg\max_i \alpha_{n,i}$ which, under the posterior, is most likely to be optimal. When the algorithm does not play $\hat{I}_n$, it plays the most likely alternative $\hat{J}_n = \arg\max_{j \neq \hat{I}_n} \alpha_{n,j}$, which is the action that is second most likely to be optimal under the posterior. Put differently, the algorithm sets $\psi_{n,\hat{I}_n} = \beta$, and $\psi_{n,\hat{J}_n} = 1 - \beta$.

3.2. Top-Two Value Sampling

We now propose a variant of top-two sampling that considers not only the probability a design is optimal, but the expected amount by which its quality exceeds that of other designs. In particular, we will define below a measure $V_{n,i}$ of the value of design i under the posterior distribution at time n. Top-two value sampling (TTVS) computes the top-two designs under this measure: $\hat{I}_n = \arg\max_i V_{n,i}$ and $\hat{J}_n = \arg\max_{j\neq \hat{I}_n} V_{n,j}$. It then plays the top design $\hat{I}_n$ with probability β and the best alternative $\hat{J}_n$ otherwise. As observations are gathered, beliefs are updated and so the top two designs change over time. The measure of value $V_{n,i}$ is defined below.

The definition of TTVS depends on a choice of (utility) function $u : \theta \mapsto \mathbb{R}$, which encodes a measure of the value of discovering a design with quality θ_i. Two natural choices of u are $u(\theta) = \theta$ and $u(\theta) = \mu(\theta)$. The paper's theoretical results allow u to be a general function, but we assume that it is *continuous* and *strictly increasing*. For a given choice of u, and any $i \in \{1, \ldots, k\}$, the function

$$\begin{aligned} v_i(\boldsymbol{\theta}) &= \max_j u(\theta_j) - \max_{j\neq i} u(\theta_j) \\ &= \begin{cases} 0 & \text{if } \boldsymbol{\theta} \notin \Theta_i \\ u(\theta_i) - \max_{j\neq i} u(\theta_j) & \text{if } \boldsymbol{\theta} \in \Theta_i \end{cases} \end{aligned}$$

provides a measure of the value of design i when the true parameter is $\boldsymbol{\theta}$. It captures the improvement in decision quality due to design i's inclusion in the choice set. Let

$$V_{n,i} = \int_{\Theta} v_i(\boldsymbol{\theta})\pi_n(\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\Theta_i} v_i(\boldsymbol{\theta})\pi_n(\boldsymbol{\theta})d\boldsymbol{\theta} \qquad (4)$$

denote the expected value of $v_i(\boldsymbol{\theta})$ under the posterior distribution at time n. This can be viewed as the option value of design i: It is the expected additional value of having the option to choose design i when it is revealed to be the best design. Note that the integral (4) defining $V_{n,i}$ is a weighted version of the integral defining $\alpha_{n,i}$. The paper will formalize a sense in which $V_{n,i}$ and $\alpha_{n,i}$ are asymptotically equivalent as $n \to \infty$, and as a result the asymptotic analysis of top-two value sampling essentially reduces to the analysis of top-two probability sampling.

3.3. Thompson Sampling

Thompson sampling is an old and popular heuristic for multiarmed problems. The algorithm simply samples actions according to the posterior probability they are optimal. In particular, it selects action i with probability $\psi_{n,i} = \alpha_{n,i}$, where $\alpha_{n,i}$ denotes the probability action i is optimal under a parameter drawn from the posterior distribution.

Thompson sampling can have very poor asymptotic performance for the best-arm identification problem. Intuitively, this is because once it estimates that a particular arm is the best with reasonably high probability, it selects that arm in almost all periods at the expense of refining its knowledge of other arms. If $\alpha_{n,i} = 0.95$, then the algorithm will only select an action other than i roughly once every 20 periods, greatly extending the time it takes until $\alpha_{n,i} > 0.99$. This insight can be made formal; our results imply that Thompson sampling attains a only attains a polynomial, rather exponential, rate of posterior convergence. A similar reasoning applies to other multiarmed bandit algorithms. The work of Bubeck et al. (2009) shows formally that algorithms satisfying regret bounds of order $\log(n)$ are necessarily far from optimal for the problem of identifying the best arm.

With this in mind, it is natural to consider a modification of Thompson sampling that simply restricts the algorithm from sampling the same action too frequently. One version of this idea is proposed below.

3.4. Top-Two Thompson Sampling

This section proposes top-two Thompson sampling (TTTS), which modifies standard Thompson sampling by adding a resampling step. As with TTPS and TTVS, this algorithm depends on a tuning parameter $\beta > 0$ that will sometimes be set to a default value of 1/2.

As in Thompson sampling, at time n, the algorithm samples a design $I \sim \boldsymbol{\alpha}_n$. Design I is measured with probability β, but, in order to prevent the algorithm from exclusively focusing on one action, with probability $1 - \beta$, an alternative design is measured. To generate this alternative, the algorithm continues sampling designs $J \sim \boldsymbol{\alpha}_n$ until the first time $J \neq I$. This can be viewed as a top-two sampling algorithm, where the top two are chosen by executing Thompson sampling until two distinct designs are drawn.

Under top-two Thompson sampling, the probability of measuring design i at time n is

$$\psi_{n,i} = \alpha_{n,i}\left(\beta + (1-\beta)\sum_{j\neq i}\frac{\alpha_{n,j}}{1-\alpha_{n,j}}\right).$$

This expression simplifies as the algorithm definitively identifies the best design. As $\alpha_{n,I^*} \to 1$, $\psi_{n,I^*} \to \beta$, and for each $i \neq I^*$,

$$\frac{\psi_{n,i}}{1-\psi_{n,I^*}} \sim \frac{\alpha_{n,i}}{1-\alpha_{n,I^*}}.$$

In this limit, the true best design is sampled with probability β. The probability i is sampled given I^* is not is equal to the posterior probability i is optimal given I^* is not.

3.5. Computing and Sampling According to Optimal Action Probabilities

Here, we provide some insight into how to efficiently implement the proposed top-two rules in important problem classes. We begin with top-two Thompson sampling, which is often the easiest to implement.

Note that given an ability to sample from Π_n, it is easy to sample from the posterior distribution over the optimal design $\boldsymbol{\alpha}_n$. In particular, if $\hat{\boldsymbol{\theta}} \sim \Pi_n$ is drawn randomly from the posterior, then $\arg\max_i \hat{\theta}_i$ is a random sample from $\boldsymbol{\alpha}_n$. Either through the choice of conjugate prior distributions or through the use of Markov chain Monte Carlo, it is possible to efficiently sample from the posterior for many interesting models. Algorithm 1 shows how to directly sample an action according to TTTS by sampling from the posterior distribution. With this algorithm, the number of times step 7 is repeated is a geometrically distributed random variable with mean $1/(1-\alpha_{n,I})$. Therefore, this step only becomes inefficient once the posterior is very highly concentrated on a single action and there is little benefit to further exploration. It is worth highlighting that this algorithm does not require computing or approximating the distribution $\boldsymbol{\alpha}_n$.

Algorithm 1 Top-Two Thompson Sampling (β)

```
1: Sample $\hat{\boldsymbol{\theta}} \sim \Pi_n$ and set $I \leftarrow \arg\max_i \hat{\theta}_i$
                                        ▷ Apply Thompson sampling
2: Sample $B \sim$ Bernoulli($\beta$)
3: if $B = 1$ then                     ▷ Occurs with probability $\beta$.
4:     Play $I$
5: else
6:     repeat
7:         Sample $\hat{\boldsymbol{\theta}} \sim \Pi_n$ and set $J \leftarrow \arg\max_j \hat{\theta}_j$
                                        ▷ Repeat Thompson sampling
8:     until $J \neq I$
9:     Play $J$
10: end if
```

The optimal action probabilities $\alpha_{n,i}$ and values $V_{n,i}$ are defined by k-dimensional integrals, which may be difficult to compute in general, even if the posterior Π_n has a closed form. Algorithm 2 shows how to approximate $\alpha_{n,i}$ and $V_{n,i}$ using samples $\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^M$, which enables efficient approximations to TTPS and TTVS whenever posterior samples can be efficiently generated.

Algorithm 2 SampleApprox(K, M, u, $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^M$)

```
$\mathcal{S}_i \leftarrow \{m | i = \arg\max_j \theta_j^m\}$    $\forall i \in \{1, \ldots, K\}$
$\hat{\alpha}_i \leftarrow |\mathcal{S}_i|/m$    $\forall i \in \{1, \ldots, K\}$
$\hat{V}_i \leftarrow M^{-1} \sum_{m \in \mathcal{S}_i} \left(u(\theta_i^m) - \max_{j \neq i} u(\theta_j^m)\right)$
        $\forall i \in \{1, \ldots, K\}$
return $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{V}}$
```

Thankfully, the computation of $\alpha_{n,i}$ and $V_{n,i}$ simplifies when the algorithm begins with an independent prior over the qualities $\theta_1, \ldots \theta_k$ of the k designs. To understand this fact, suppose $X_1, \ldots, X_k \in \mathbb{R}$ are independently distributed and continuous random variables. Then,

$$\mathbb{P}\left(X_1 = \max_i X_i\right) = \int_{x \in \mathbb{R}} f_1(x) \prod_{j=2}^{k} F_j(x) dx, \tag{5}$$

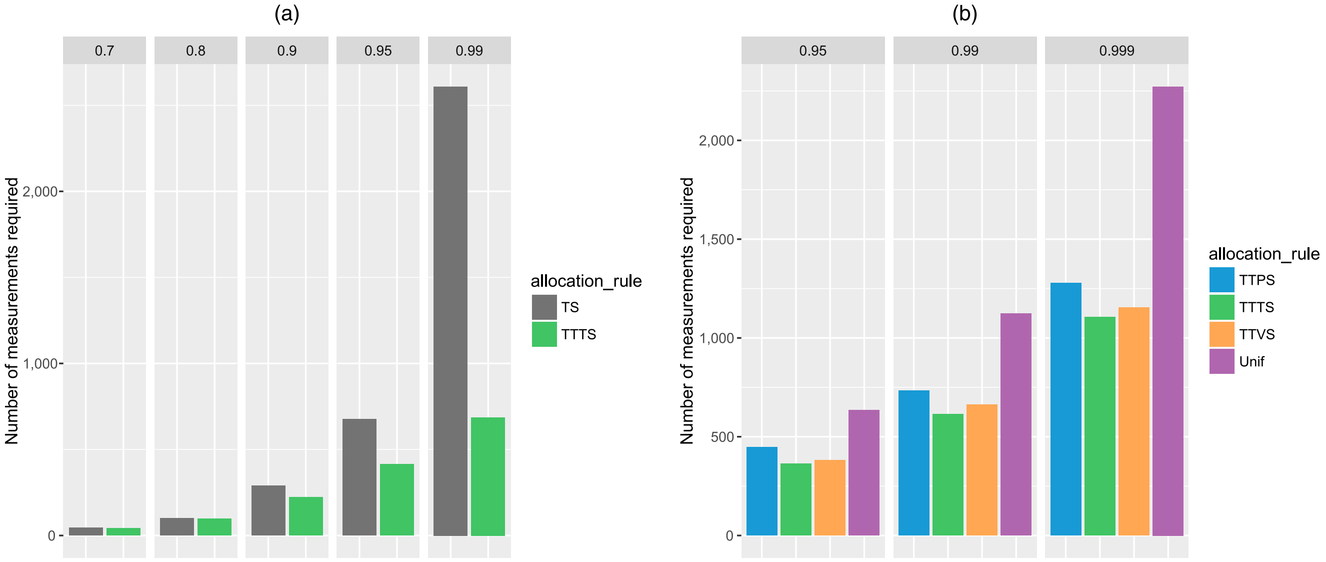
where f_1 denotes the probability density function of X_1 and $F_2, \ldots, F_K$ are the cumulative distribution functions of $X_2, \ldots, X_k$. In particular, $\mathbb{P}(X_1 = \max_i X_i)$ can be computed by solving a one-dimensional integral. Based on this insight, Online Appendix EC.3 provides an efficient implementation of TTPS for a problem with independent beta priors and binary observations. That implementation approximates one-dimensional integrals like (5) using quadrature with m points and has the time and space complexity that scale as $O(km)$.

4. First Insights from a Numerical Experiment

Some of the paper's main insights are reflected in a simple numerical experiment. Consider a problem where observations are binary $Y_{n,i} \in \{0, 1\}$, and the unknown vector $\boldsymbol{\theta}^* = (.1, .2, .3, .4, .5)$ defines the true success probability of each design. Each algorithm begins with an independent uniform prior over the components of $\boldsymbol{\theta}^*$. The experiment compares the performance of top-two probability sampling, top-two value sampling,[2] and top-two Thompson sampling with $\beta = 1/2$ against Thompson sampling and a uniform allocation rule, which allocates equal measurement effort ($\psi_{n,i} = 1/5$) to each design. The uniform allocation is a natural point of comparison. In the author's experience, it is widely used in practice in a variety of domains. Figure 1 displays the average number of measurements required for the posterior to reach a given confidence level. In particular, the experiment tracks the first time when $\max_i \alpha_{n,i} \geq c$ for various confidence levels $c \in (0, 1)$. Figure 1 displays the average number of measurements required for each algorithm to reach each fixed confidence level, where the average was taken over 100 trials in panel (a) and 500 in panel (b). Even for this simple problem with five designs, the proposed algorithms can reach the same confidence level by using fewer than half the measurements required by a uniform allocation rule. Although all the top-two rules attain the same asymptotic rate of convergence, we can see that top-two probability sampling is slightly outperformed in this experiment. Figure 1(a) compares Thompson sampling to top-two Thompson sampling. Thompson sampling (TS) appears to reach low confidence levels as rapidly as top-two TS, but as suggested in Section 3.3, is very slow to reach high levels of confidence. It requires more than 60% more measurements to reach confidence 0.95 and over 250% more measurements to reach confidence 0.99. TS requires an onerous number of measurements to reach confidence 0.999, and so we omit this experiment.

Figure 2 provides insight into how the proposed algorithms differ from the uniform allocation. It displays the distribution of measurements and posterior

Figure 1. (Color online) Number of Measurements Required to Reach Given Confidence Level

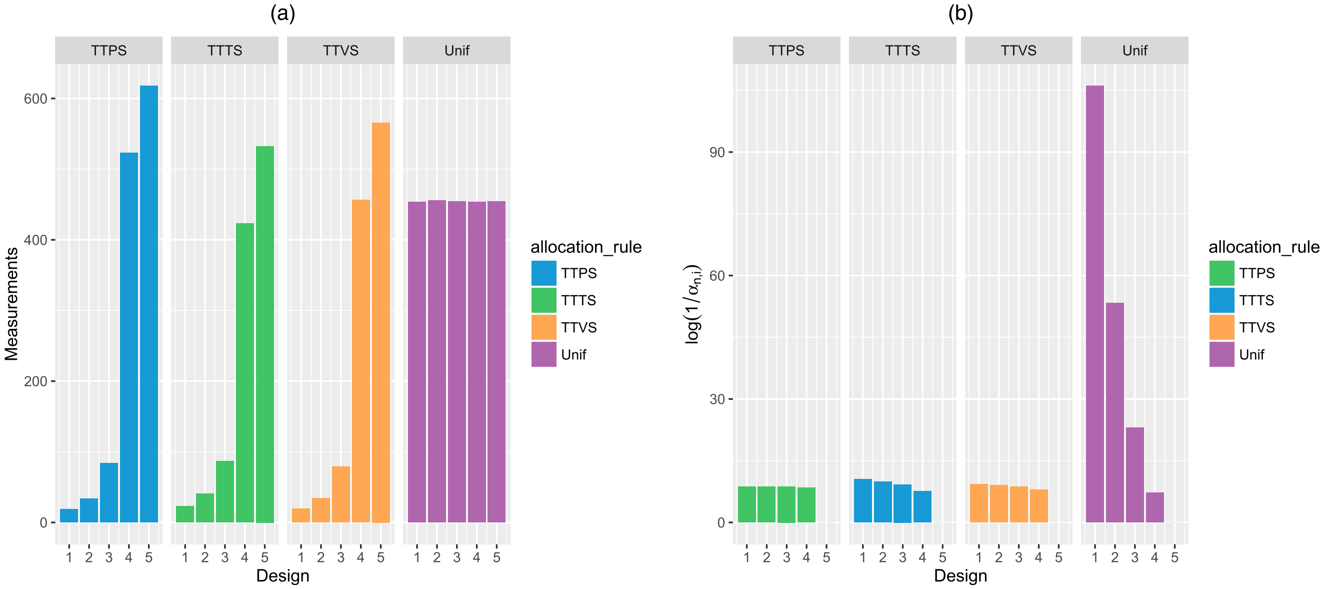


Notes. (a) TS vs. top-two TS. (b) Comparison with uniform allocation.

beliefs at the first time when a confidence level of 0.999 is reached. Again, all results are averaged across 500 trials. Figure 2(a) displays the average number of measurements collected from each design. It is striking that although TTTS, TTPS, and TTVS seem quite different, they all settle on essentially the same distribution of measurement effort. Because $\beta = 1/2$, roughly one-half of the measurements are collected from $I^* = 5$. Moreover, fewer measurements are collected from designs that are farther from optimal, and most of the remaining half of measurement effort is allocated to design 4. Notice that using the same number of noisy samples, it is much more difficult certify that $\theta^*_4 < \theta^*_5$ than that $\theta^*_1 < \theta^*_5$, both because θ^*_4 is closer to θ^*_5 and because observations from a Bernoulli distribution with parameter 0.4 have higher variance than under a Bernoulli distribution with parameter 0.1.

Figure 2(b) investigates the posterior probability $\alpha_{n,i}$ assigned to the event that design i is optimal. To make the insights more transparent, these are plotted on log-scale, where the value $\log(1/\alpha_{n,i})$ can roughly be interpreted as the magnitude of evidence that alternative i is not optimal. By using an *equal allocation* of

Figure 2. (Color online) Distribution of Measurements and Posterior Beliefs at Termination



Notes. (a) Measurements collected of each design. (b) Value of $\log(1/\alpha_{n,i})$ for each design i.

measurement effort across the designs, the uniform sampling rule gathers an enormous amount of evidence to rule out design 1, but an order-of-magnitude less evidence to rule out design 4. Instead of allocating measurement effort equally across the alternatives, TTTS, TTPS, and TTVS appear to exactly adjust measurement effort to gather *equal evidence* that each of the first four designs is not optimal.

Intuitively, in the long run, each of the proposed algorithms will allocate measurement effort to design 5—the true best design—and to whichever other designs could most plausibly be optimal. If too much measurement effort has been allocated to a particular design, then the posterior will indicate that it is clearly suboptimal, and effort will be allocated elsewhere until a similar amount of evidence has been gathered about other designs. In this way, measurement effort is automatically adjusted to the appropriate level.

5. Main Theoretical Results

Our main theoretical results concern the frequentist consistency and rate of convergence of the posterior distribution. Recall that

$$\Pi_n(\Theta^c_{I^*}) = \sum_{i \neq I^*} \alpha_{n,i}$$

captures the posterior mass assigned to the event that an action other than I^* is optimal. One hopes that $\Pi_n(\Theta^c_{I^*}) \to 0$ as the number of observations n tends to infinity, so that the posterior distribution converges on the truth. We will show that under the TTTS, TTPS, and TTVS allocation rules, $\Pi_n(\Theta^c_{I^*})$ converges to zero an exponential rate and that the exponent governing the rate of convergence is nearly the best possible.

To facilitate theoretical analysis, we will make three additional boundedness assumptions, which are assumed throughout all formal proofs. This rules out some cases of interest, such as the use of multivariate Gaussian priors. However, we otherwise allow for quite general correlated priors, expressed in terms of a density over a compact set. This stands in contrast, for example, to previous analyses of Thompson sampling, which typically rely heavily on the use of independent conjugate priors. Assumption 1 is used only in establishing certain asymptotic results concerning the rate of posterior concentration. Analogous results are easily established for certain unbounded conjugate priors,[3] but the author still has not identified the right technical conditions that generalize these results.

Assumption 1. *The parameter space is a bounded open hyper-rectangle $\Theta = (\underline{\theta}, \overline{\theta})^k$, the prior density is uniformly bounded with*

$$0 < \inf_{\boldsymbol{\theta} \in \Theta} \pi_1(\boldsymbol{\theta}) < \sup_{\boldsymbol{\theta} \in \Theta} \pi_1(\boldsymbol{\theta}) < \infty,$$

and the log-partition function has bounded first derivative with $\sup_{\theta \in [\underline{\theta}, \overline{\theta}]} |A'(\theta)| < \infty$.

The paper's main results, as stated in the next theorem, characterize the rate of posterior convergence under the proposed algorithms, formalize a sense in which this is the fastest possible rate, and bound the impact of the tuning parameter $\beta \in (0,1)$. The statement depends on distribution-dependent constants $\Gamma^*_\beta > 0$ and $\Gamma^* > 0$ that are presented here but will be more explicitly characterized in Section 6.

The first part of the theorem shows that there is an exponent $\Gamma^* > 0$ such that $\Pi_n(\Theta^c_{I^*})$ cannot converge to zero at a rate faster than $e^{-n\Gamma^*}$ under any allocation rule and shows that TTPS, TTVS, and TTTS attain this optimal rate of convergence when the tuning parameter β is set optimally. This optimal exponent is shown to equal

$$\Gamma^* = \max_{\psi} \min_{\boldsymbol{\theta} \in \Theta^c_{I^*}} \sum_{i=1}^{k} \psi_i d(\theta^*_i \| \theta_i),$$

where $d(\theta_i \| \theta'_i)$ denotes the Kullback–Leibler divergence between the observation distributions $p(y|\theta_i)$ and $p(y|\theta'_i)$. This can be viewed as the value of a game between two players. An experimenter first chooses a probability distribution $\boldsymbol{\psi}$, determining the frequency with which arms are measured. An adversary then chooses the worst-case configuration of arm means, selecting an alternative $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ that is hard to distinguish from $\boldsymbol{\theta}^*$ under the measurement allocation $\boldsymbol{\psi}$, but under which the arm I^* is no longer optimal. Complexity terms of this form date back over 60 years to classic work of Chernoff (1959) on the sequential design of experiments.

The remainder of the theorem investigates the role of the tuning parameter $\beta \in (0,1)$. Part 2 shows that there is an exponent $\Gamma^*_\beta > 0$ such that $\Pi_n(\Theta^c_{I^*}) \to 0$ at rate $e^{-n\Gamma^*_\beta}$ under TTPS, TTVS, or TTTS with parameter β, and this is shown to be optimal among a restricted class of allocation rules. In particular, we observe that β controls the fraction of measurement effort allocated to the true best design I^*, in the sense that $\overline{\psi}_{n,I^*} \to \beta$ as $n \to \infty$ under each of the proposed algorithms. These algorithms attain the error exponent

$$\Gamma^*_\beta = \max_{\psi: \psi_{I^*} = \beta} \min_{\boldsymbol{\theta} \in \Theta^c_{I^*}} \sum_{i=1}^{k} \psi_i d(\theta^*_i \| \theta_i),$$

which differs from Γ^* because the experimenter is constrained to measure the true best arm with fraction β of measurement effort. A lower bound shows that this exponent is optimal among a constrained class: Precisely, on any sample path on which an adaptive algorithm allocates a faction β of overall effort to measuring I^*, the posterior cannot converge at rate faster than $e^{-n\Gamma^*_\beta}$. In this sense, while a tuning parameter controls the long-run measurement effort allocated to the true best design, TTPS, TTVS, and TTTS all automatically adjust how the remaining measurement effort is allocated among the $k-1$ suboptimal designs in an asymptotically optimal manner.

The final part of the theorem shows that the constrained exponent Γ^*_β is close to the largest possible exponent Γ^* whenever β is close to the optimal value. The choice of $\beta = 1/2$ is particularly robust: $\Gamma^*_{1/2}$ is never more than a factor of 2 away from the optimal exponent. This result implies that $e^{-2n\Gamma^*_{1/2}} \le e^{-n\Gamma^*}$, so $2n$ samples collected with top-two Thompson sampling using $\beta = 1/2$ give a faster asymptotic rate of posterior convergence than n samples collected using any adaptive algorithm.

Theorem 1. *There exist constants $\{\Gamma^*_\beta > 0 : \beta \in (0,1)\}$ such that $\Gamma^* = \max_\beta \Gamma^*_\beta$ exists, $\beta^* = \arg\max_\beta \Gamma^*_\beta$ is unique, and the following properties are satisfied with probability 1:*

1. *Under TTTS, TTPS, or TTVS with parameter β^*,*

$$\lim_{n\to\infty} -\frac{1}{n}\log \Pi_n(\Theta^c_{I^*}) = \Gamma^*.$$

Under any adaptive allocation rule,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \Pi_n(\Theta^c_{I^*}) \le \Gamma^*.$$

2. *Under TTTS, TTPS, or TTVS with parameter $\beta \in (0,1)$,*

$$\lim_{n\to\infty} -\frac{1}{n}\log \Pi_n(\Theta^c_{I^*}) = \Gamma^*_\beta \quad \text{and} \quad \lim_{n\to\infty} \overline{\psi}_{n,I^*} = \beta.$$

Under any adaptive allocation rule,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \Pi_n(\Theta^c_{I^*}) \le \Gamma^*_\beta$$

$$\text{on any sample path with} \quad \lim_{n\to\infty} \overline{\psi}_{n,I^*} = \beta.$$

3. $\Gamma^* \le 2\Gamma^*_{\frac{1}{2}}$ *and*

$$\frac{\Gamma^*}{\Gamma^*_\beta} \le \max\left\{\frac{\beta^*}{\beta}, \frac{1-\beta^*}{1-\beta}\right\}.$$

This theorem is established in a sequence of results in Section 6. The lower bounds in parts 1 and 2 are given, respectively, in Propositions 6 and 7. Proposition 8 shows that the top-two rules attain these optimal exponents. Part 3 is stated as Lemma 3 in Section 6.

A key element of the proof is to show that, under the proposed top-two sampling algorithms applied to any problem instance $\boldsymbol{\theta}^*$, the long-run fraction of samples collected from each arm converges almost surely to the vector $\boldsymbol{\psi}^*_\beta(\boldsymbol{\theta}^*)$ that attains the maximum in the definition of Γ^*_β. If β is set or tuned appropriately, sampling proportions converge almost surely to the sampling proportions that attain the maximum in the definition of Γ^*. Interestingly, these optimal asymptotic sampling proportions and the exponent Γ^* have been derived several times, seemingly independently, by authors aiming to optimize different performance criteria (Chernoff 1959, Jennison et al. 1982, Glynn and Juneja 2004). In this sense, the allocation $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$ to which top-two algorithms converge appears to be linked to many notions of optimal performance in best-arm identification problems. However, substantial subtleties arise because convergence to this allocation with probability 1, as established in this paper, may not be sufficient alone to guarantee optimality according to these performance criteria. A more complete discussion of related work is given in Section 1.2, and a more technical discussion of alternative performance criteria is given in Online Appendix EC.2.

5.1. An Upper Bound on the Error Exponent

Before proceeding, we will state an upper bound on the error exponent when $\beta = 1/2$ that is closely related to complexity terms that have appeared in the literature on best-arm identification (e.g., Audibert et al. 2010). This bound depends on the gaps between the means of the different observation distributions.

We say that a real valued random variable X is σ-sub-Gaussian if $\mathbb{E}\big[\exp\{\lambda(X-\mathbb{E}[X])\}\big] \le \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\}$ so that the moment-generating function of $X-\mathbb{E}[X]$ is dominated by that of a zero mean Gaussian random variable with variance σ^2. Gaussian random variables are sub-Gaussian, as are uniformly bounded random variables. The next result applies to both Bernoulli and Gaussian distributions, as each can be parameterized with sufficient statistic $T(y) = y$.

Proposition 1. *Suppose the exponential family distribution is parameterized with $T(y) = y$ and that for each $\theta \in [\underline{\theta}, \overline{\theta}]$, if $Y \sim p(y|\theta)$, then Y is sub-Gaussian with parameter σ. Then,*

$$\Gamma^*_{\frac{1}{2}} \ge \frac{1}{16\sigma^2 \sum_{i\neq I^*} \Delta_i^{-2}},$$

where for each $i \in \{1,\ldots,k\}$,

$$\Delta_i = \mathbb{E}[Y_{n,I^*}] - \mathbb{E}[Y_{n,i}]$$

is the difference between the mean under $\theta^*_{I^*}$ and the mean under θ^*_i.

This shows that $\Pi_n(\Theta^c_{I^*})$ decays at asymptotic rate faster than $\exp\left\{-\frac{n\min_i \Delta_i^2}{16k\sigma^2}\right\}$, so convergence is rapid when there is a large gap between the means of different designs. In fact, Proposition 1 replaces the dependence on $(1/k)$ times the smallest gap Δ_i with a dependence on $(\sum_{i=2}^k \Delta_i^{-2})^{-1}$, which captures the average inverse gap. This rate is attained only by an intelligent adaptive algorithm that allocates more measurement effort to designs that are nearly optimal and less to designs that are clearly suboptimal. In fact, the next result shows that the asymptotic performance of uniform allocation rule depends only on the smallest gap $\min_{i\neq I^*}\Delta_i^2$, and therefore even if some designs could be quickly ruled out, the algorithm can't leverage this to attain a faster rate of convergence.

Proposition 2. *If $Y_{n,I^*}\sim\mathcal{N}(0,\sigma^2)$ and $Y_{n,i}\sim\mathcal{N}(-\Delta_i,\sigma^2)$ for each $i\neq I^*$,*

$$\lim_{n\to\infty} -\frac{1}{n}\log\Pi_n\left(\Theta_{I^*}^c\right)=\frac{-\min_i\Delta_i^2}{4k\sigma^2}$$

under a uniform allocation rule, which sets $\psi_{n,i}=1/k$ for each i and n.

5.2. Consistent Tuning of β

Our previous results show that if the top-two sampling algorithms are applied with the optimal problem-dependent tuning parameter $\beta^*=\arg\max_\beta\Gamma^*$, then these algorithms attain the optimal rate of posterior convergence $e^{-\Gamma^* n}$. Unfortunately, β^* is typically unknown, and so we also investigate robustness to the choice of β, both in theory, as in Theorem 1 above, and in simulation experiments presented in Section 7. Still, a natural question is whether this tuning parameter can be adjusted in a dynamic fashion to converge on β^*. We initiate the study of such extensions in this section.

First, note that it is easy to extend the definition of each top-two sampling algorithm so that they use an adaptive sequence of tuning parameters $(\beta_n : n\in\mathbb{N})$. For example, top-two probability sampling identifies $\hat{I}_n=\arg\max_i\alpha_{n,i}$ and $\hat{J}_n=\arg\max_{j\neq\hat{I}_n}\alpha_{n,j}$ and then chooses among these with respective probabilities $\psi_{n,\hat{I}_n}=\beta_n$ and $\psi_{n,\hat{J}_n}=1-\beta_n$. The next lemma confirms that, if applied with such a sequence of tuning parameters such that $\beta_n\to\beta^*$, the top-two sampling algorithms attain the optimal convergence rate $e^{-n\Gamma_{\beta^*}}$.

Proposition 3. *Suppose TTTS, TTVS, and TTPS are applied with an adaptive sequence of tuning parameters $(\beta_n : n\in\mathbb{N})$, where for each n, β_n is $\mathcal{F}_{n-1}$ measurable. Then, with probability 1, on any sample path on which $\beta_n\to\beta^*$,*

$$\Pi_n\left(\Theta_{I^*}^c\right)\doteq e^{-n\Gamma^*}.$$

The next lemma confirms that such consistent tuning is possible. The method for tuning β, presented in Algorithm 3, simply solves numerically for the optimal value of β assuming that the true values of the parameters $(\theta_1,\ldots\theta_k)$ are given by their respective posterior means.

Unfortunately, this tuning method is complex, spoiling some of elegance of the top-two sampling algorithms. A significant open question is whether simpler methods for adapting β could be adopted.

Lemma 1. *Under TTTS, TTPS, or TTVS with an adaptive sequence of tuning parameters $(\beta_n : n\in\mathbb{N})$ adjusted according to Algorithm 3, $\beta_n\to\beta^*$ almost surely. Therefore, $\Pi_n(\Theta_{I^*}^c)\doteq e^{-n\Gamma^*}$.*

Algorithm 3 Top-Two Sampling with β-Tuning

```
Input $\kappa\geq 2$, $\hat{\beta}\in(0,1)$.
Set counter $\ell=1$
for $n\in\{1,2,3,4,\ldots\}$ do
    Sample $I_n\sim\mathrm{TopTwo}(\pi_n,\hat{\beta})$
    Measure $I_n$ and observe $Y_{n,I_n}$
    Update play-count $S_{n+1,I_n}\leftarrow S_{n,I_n}+1$
    Update posterior $\pi_{n+1}(\boldsymbol{\theta})\propto\pi_n(\boldsymbol{\theta})p(Y_{n,I_n}\mid\theta_{I_n})$
    if $\min_i S_{n,i}\geq\kappa^\ell$ then
        $\ell\leftarrow\ell+1$
        Compute posterior mean $\hat{\boldsymbol{\theta}}\leftarrow\int_\Theta\boldsymbol{\theta}\pi_{n+1}(\boldsymbol{\theta})d\boldsymbol{\theta}$
        if $\arg\max_i\hat{\theta}_i$ is unique then
            Estimate best arm $\hat{I}\leftarrow\arg\max_i\hat{\theta}_i$
            Estimate best allocation
                $\hat{\boldsymbol{\psi}}\leftarrow\arg\max_{\boldsymbol{\psi}}\min_{\boldsymbol{\theta}\in\Theta_{\hat{I}}^c}D_\psi(\hat{\boldsymbol{\theta}}\|\boldsymbol{\theta})$
            $\hat{\beta}\leftarrow\hat{\psi}_{\hat{I}}$
        end if
    end if
end for
```

6. Analysis

6.1. Asymptotic Notation

To simplify the presentation, it is helpful to introduce additional asymptotic notation. We say two sequences a_n and b_n taking values in $\mathbb{R}$ are *logarithmically equivalent*, denoted by $a_n\doteq b_n$, if $\frac{1}{n}\log\left(\frac{a_n}{b_n}\right)\to 0$ as $n\to\infty$. This notation means that a_n and b_n are *equal up to first order in the exponent*. With this notation, Theorem 1 implies that the top-two sampling rules with parameter β attain the convergence rate $\Pi_n(\Theta_{I^*}^c)\doteq e^{-n\Gamma_\beta^*}$. This is an equivalence relation, in the sense that if $a_n\doteq b_n$ and $b_n\doteq c_n$, then $a_n\doteq c_n$. Note that $a_n+b_n\doteq\max\{a_n,b_n\}$, so that the sequence with the largest exponent dominates. In addition, for any positive constant c, $ca_n\doteq a_n$, so that constant multiples of sequences are equal up to first order in the exponent. When applied to sequences of random variables, these relations are understood to apply almost surely.

It is natural to wonder whether the proposed algorithms asymptotically minimize expressions like $\sum_{i\neq I^*}(\theta_{I^*}^*-\theta_i)\alpha_{n,i}$, which account for how far some designs are from optimal. We note, in passing, that

$$\sum_{i\neq I^*}c_i\alpha_{n,i}\doteq\max_{i\neq I^*}\alpha_{n,i}$$

for any positive costs $c_i>0$, and so any such performance measures are equal to first order in the exponent. Similar observations have been used to justify the study of the probability of incorrect selection, rather than notions of the expected cost of an incorrect decision (Glynn and Juneja 2004, Audibert et al. 2010).

6.2. Posterior Consistency

The next proposition provides a consistency and anticonsistency result for the posterior distribution.

The first part says that if design i receives infinite measurement effort, the marginal posterior distribution of its quality concentrates around the true value θ_i^*. The second part says that when restricted to designs that are not measured infinitely often, the posterior does not concentrate around any value. The posterior converges to the truth as infinite evidence is collected, but nothing can be ruled out with certainty based on finite evidence.

Proposition 4. *With probability* 1, *for any* $i \in \{1,..,k\}$, *if* $\Psi_{n,i} \to \infty$, *then, for all* $\epsilon > 0$,

$$\Pi_n\left(\left\{\boldsymbol{\theta} \in \Theta \middle| \theta_i \notin \left(\theta_i^* - \epsilon, \theta_i^* + \epsilon\right)\right\}\right) \to 0.$$

If $\mathcal{I} = \{j \in \{1,\dots,k\} \mid \lim_{n\to\infty} \Psi_{n,j} < \infty\}$ *is nonempty, then*

$$\inf_{n\in\mathbb{N}} \Pi_n\left(\left\{\boldsymbol{\theta} \in \Theta \middle| \theta_i \in \left(\theta_i', \theta_i''\right) \; \forall i \in \mathcal{I}\right\}\right) > 0$$

for any collections of open intervals $(\theta_i', \theta_i'') \subset (\underline{\theta}, \overline{\theta})$ *ranging over* $i \in \mathcal{I}$.

This result is the key to establishing that $\alpha_{n,I^*} \to 1$ under each of the proposed algorithms. The next subsection gives a more refined result that allows us to characterize the rate of convergence.

6.3. Posterior Large Deviations

This section provides an asymptotic characterization of posterior probabilities $\Pi_n(\tilde{\Theta})$ for any open set $\tilde{\Theta} \subset \Theta$ and under any adaptive measurement strategy. The characterization depends on the notion of Kullback–Leibler divergence. For two parameters $\theta, \theta' \in \mathbb{R}$, the log-likelihood ratio, $\log(p(y|\theta)/p(y|\theta'))$, provides a measure of the amount of information y provides in favor of θ over θ'. The Kullback–Leibler divergence

$$d(\theta\|\theta') \triangleq \int \log\left(\frac{p(y|\theta)}{p(y|\theta')}\right) p(y|\theta) d\nu(y)$$

is the expected value of the log-likelihood under observations drawn $p(y|\theta)$. Then, if the design to measure is chosen by sampling from a probability distribution $\boldsymbol{\psi}$ over $\{1,..,k\}$,

$$D_{\psi}(\boldsymbol{\theta}\|\boldsymbol{\theta}') \triangleq \sum_{i=1}^{k} \psi_i d(\theta_i\|\theta_i')$$

is the average Kullback–Leibler divergence between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ under $\boldsymbol{\psi}$.

Under the algorithms, we consider, the effort allocated to measuring design i, $\psi_{n,i} \triangleq \mathbb{P}(I_n = i|\mathcal{F}_{n-1})$, changes over time as data are collected. Recall that $\overline{\boldsymbol{\psi}}_{n,i} \triangleq n^{-1} \sum_{\ell=1}^{n} \psi_{\ell,i}$ captures the fraction of overall effort allocated to measuring design i over the first n periods. Under an adaptive allocation rule, $\overline{\boldsymbol{\psi}}_n$ is function of the history $(I_1, Y_{1,I_1}, \dots I_{n-1}, Y_{n-1,I_{n-1}})$ and is therefore a random variable. Given that measurement effort has been allocated according to $\overline{\boldsymbol{\psi}}_n$, $D_{\overline{\psi}_n}(\boldsymbol{\theta}^*\|\boldsymbol{\theta})$ quantifies the average information acquired that distinguishes $\boldsymbol{\theta}$ from the true parameter $\boldsymbol{\theta}^*$. The following proposition relates the posterior mass assigned to $\tilde{\Theta}$ to $\inf_{\theta\in\tilde{\Theta}} D_{\overline{\psi}_n}(\boldsymbol{\theta}^*\|\boldsymbol{\theta})$, which captures the element in $\tilde{\Theta}$ that is hardest to distinguish from $\boldsymbol{\theta}^*$ based on samples from $\overline{\boldsymbol{\psi}}_n$.

Proposition 5. *For any open set* $\tilde{\Theta} \subset \Theta$,

$$\Pi_n(\tilde{\Theta}) \doteq \exp\left\{-n \inf_{\boldsymbol{\theta}\in\tilde{\Theta}} D_{\overline{\psi}_n}(\boldsymbol{\theta}^*\|\boldsymbol{\theta})\right\}.$$

To understand this result, consider a simpler setting where the algorithm measures design i in every period, and consider some $\boldsymbol{\theta}$ with $\theta_i \neq \theta_i^*$. Then, the log-ratio of posterior densities

$$\log\left(\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)}\right) = \log\left(\frac{\pi_1(\boldsymbol{\theta})}{\pi_1(\boldsymbol{\theta}^*)}\right) + \sum_{\ell=1}^{n-1} \log\left(\frac{p(Y_{\ell,i}|\theta_i)}{p(Y_{\ell,i}|\theta_i^*)}\right)$$

can be written as the sum of the log-prior ratio and the log-likelihood ratio. The log-likelihood ratio is negative drift random walk: It is the sum of $n-1$ independent and identically distributed terms, each of which has mean

$$\mathbb{E}\left[\log\left(\frac{p(Y_{1,i}|\theta_i)}{p(Y_{1,i}|\theta_i^*)}\right)\right] = \mathbb{E}\left[-\log\left(\frac{p(Y_{1,i}|\theta_i^*)}{p(Y_{1,i}|\theta_i)}\right)\right]$$
$$= -d(\theta_i^*\|\theta_i).$$

Therefore, by the law of large numbers, as $n \to \infty$, $n^{-1}\log(\pi_n(\boldsymbol{\theta})/\pi_n(\boldsymbol{\theta}^*)) \to -d(\theta_i^*\|\theta_i)$, or, equivalently, the ratio of the posterior densities decays exponentially as

$$\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} \doteq \exp\{-nd(\theta_i^*\|\theta_i)\}.$$

This calculation can be carried further to show that if the designs measured $(I_1, I_2, I_3, \dots)$ are drawn independently of the observations $(\boldsymbol{Y_1}, \boldsymbol{Y_2}, \boldsymbol{Y_3}, \dots)$ from a fixed probability distribution $\boldsymbol{\psi}$, then

$$\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} \doteq \exp\{-nD_{\psi}(\boldsymbol{\theta}^*\|\boldsymbol{\theta})\}. \tag{6}$$

Now, by a Laplace approximation, one might expect that the integral $\int_{\tilde{\Theta}} \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta}$ is extremely well approximated by integrating around a vanishingly small ball around the point

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\tilde{\Theta}} D_{\psi}(\boldsymbol{\theta}^*\|\boldsymbol{\theta}).$$

These are the main ideas behind Proposition 5, but there are several additional technical challenges involved in a rigorous proof. First, we need that a property like (6) holds when the allocation rule is adaptive to the data. Next, convergence of the integral of the posterior density requires a form of uniform convergence in (6). Finally, since $\overline{\psi}_n$ changes over time, the point $\arg\min_{\boldsymbol{\theta}\in\tilde{\Theta}} D_{\overline{\psi}_n}(\boldsymbol{\theta}^*\|\boldsymbol{\theta})$ changes over time, and basic Laplace approximations don't directly apply.

6.4. Characterizing the Optimal Allocation

Throughout this paper, an experimenter wants to gather enough evidence to certify that I^* is optimal, but because she does not know $\boldsymbol{\theta}^*$, she does not know which measurements will provide the most information. To characterize the optimal exponent Γ^*, however, it is useful to consider the easier problem of gathering the most effective evidence when $\boldsymbol{\theta}^*$ is known. We can cast this as a game between two players:

- An experimenter, who knows the true parameter $\boldsymbol{\theta}^*$, chooses a (possibly adaptive) measurement rule.
- A referee observes the resulting sequence of observations $(I_1, Y_{1,I_1}, \ldots, I_{n-1}, Y_{n-1,I_{n-1}})$ and computes posterior beliefs $(\alpha_{n,1}, \ldots, \alpha_{n,k})$ according to Bayes rule (2, 3).
- How can the experimenter gather the most compelling evidence? A rule which is optimal asymptotically should maximize the rate at which $\alpha_{n,I^*} \to 1$ as $n \to \infty$.

In order to drive the posterior probability α_{n,I^*} to 1, the decision maker must be able to rule out all parameters in $\Theta^c_{I^*}$ under which the optimal action is not I^*. Our analysis shows that the posterior probability assigned to $\Theta^c_{I^*}$ is dominated by the parameter that is hardest to distinguish from $\boldsymbol{\theta}^*$ under $\overline{\psi}_n$. In particular, by Proposition 5,

$$\Pi_n(\Theta^c_{I^*}) \doteq \exp\left\{-n\left(\min_{\boldsymbol{\theta}\in\Theta^c_{I^*}} D_{\overline{\psi}_n}(\boldsymbol{\theta}^*\|\boldsymbol{\theta})\right)\right\},$$

as $n \to \infty$. Therefore, the solution to the max-min problem

$$\max_{\psi} \min_{\boldsymbol{\theta}\in\Theta^c_{I^*}} D_{\psi}(\boldsymbol{\theta}^*\|\boldsymbol{\theta}), \tag{7}$$

represents an asymptotically optimal allocation rule. As highlighted in the literature review, the max-min problem (7) closely mirrors the main sample complexity term in Chernoff's classic paper on the sequential design of experiments (Chernoff 1959).

Simplifying the Optimal Exponent. Thankfully, the best-arm identification problem has additional structure that allows us to simplify the optimization problem (7). Much of our analysis involves the posterior probability assigned to the event some action $i \neq I^*$ is optimal. This can be difficult to evaluate, because the set of parameter vectors under which i is optimal

$$\Theta_i = \{\boldsymbol{\theta} \in \Theta | \theta_i \geq \theta_1, \ldots \theta_i \geq \theta_k\},$$

involves k separate constraints. Consider instead a simpler problem of comparing the parameter θ^*_i against $\theta^*_{I^*}$. For each $i \neq I^*$, define the set

$$\overline{\Theta}_i \triangleq \{\boldsymbol{\theta} \in \Theta | \theta_i \geq \theta_{I^*}\} \supset \Theta_i,$$

under which the value at i exceeds that at I^*. Because, ignoring the boundary of the set, $\Theta^c_{I^*} = \cup_{i\neq I^*}\overline{\Theta}_i$,

$$\max_{i\neq I^*} \Pi_n(\overline{\Theta}_i) \leq \Pi_n(\Theta^c_{I^*}) \leq k \max_{i\neq I^*} \Pi_n(\overline{\Theta}_i),$$

and, therefore,

$$\Pi_n(\Theta^c_{I^*}) \doteq \max_{i\neq I^*} \Pi_n(\overline{\Theta}_i). \tag{8}$$

This yields an analogue of (7) that will simplify our subsequent analysis. Combining (8) with Proposition 5 shows that the solution to the max-min problem

$$\Gamma^* \triangleq \max_{\psi} \min_{i\neq I^*} \min_{\boldsymbol{\theta}\in\overline{\Theta}_i} D_{\psi}(\boldsymbol{\theta}^*\|\boldsymbol{\theta}) \tag{9}$$

represents an asymptotically optimal allocation rule. Because the set $\overline{\Theta}_i$ involves only a constraints on θ_i and θ_{I^*}, we can derive an expression of the inner-minimization problem over $\boldsymbol{\theta}$ in terms of the measurement effort allocated to i and I^*. Define

$$C_i(\beta, \psi) \triangleq \min_{x\in\mathbb{R}} \beta d(\theta^*_{I^*}\|x) + \psi d(\theta^*_i\|x). \tag{10}$$

The next lemma shows that the function C_i arises as the solution to the minimization problem over $\boldsymbol{\theta} \in \overline{\Theta}_i$ in (9). It also shows that the minimum in (10) is attained by a parameter $\overline{\theta}$, under which the mean observation is a weighted combination of the means under $\theta^*_{I^*}$ and θ^*_i. Recall that, for an exponential family distribution $A'(\theta) = \int T(y)p(y|\theta)d\nu(y)$ is the mean observation of the sufficient statistic $T(y)$ under θ.

Lemma 2. *For any $i \in \{1, .., k\}$ and probability distribution $\boldsymbol{\psi}$ over $\{1, \ldots, k\}$,*

$$\min_{\boldsymbol{\theta}\in\overline{\Theta}_i} D_{\psi}(\boldsymbol{\theta}^*\|\boldsymbol{\theta}) = C_i(\psi_{I^*}, \psi_i).$$

In addition, each C_i is a strictly increasing concave function satisfying

$$C_i(\psi_{I^*}, \psi_i) = \psi_{I^*} d(\theta^*_{I^*}\|\overline{\theta}) + \psi_i d(\theta^*_i\|\overline{\theta}),$$

where $\overline{\theta} \in [\theta^*_i, \theta^*_{I^*}]$ is the unique solution to

$$A'(\overline{\theta}) = \frac{\psi_{I^*} A'(\theta^*_{I^*}) + \psi_i A'(\theta^*_i)}{\psi_{I^*} + \psi_i}.$$

Lemma 2 and Equation (9) immediately imply

$$\Gamma^* = \max_{\psi} \min_{i\neq I^*} C_i(\psi_{I^*}, \psi_i). \tag{11}$$

This result essentially simplifies the earlier form of the exponent, which is similar to a problem complexity measure in Chernoff (1959), into a form that mirrors[4] the large deviations exponent suggested in Glynn and Juneja (2004). The function $C_i(\beta,\psi)$ captures the effectiveness with which one can certify $\theta^*_{I^*} \geq \theta^*_i$ using an allocation rule that measures actions I^* and i with respective frequencies β and ψ. Naturally, it is an increasing function of the measurement effort (β,ψ) allocated to designs I^* and i. For given β and ψ, $C_i(\beta,\psi) \geq C_j(\beta,\psi)$ when $\theta^*_i \leq \theta^*_j$, reflecting that θ^*_i is easier to distinguish from $\theta^*_{I^*}$ than θ^*_j.

Example 1 (Gaussian Observations)**.** Suppose each outcome distribution $p(y|\theta^*_i)$ is Gaussian with unknown mean θ^*_i. Then, direct calculation using Lemma 2 shows

$$C_i(\beta,\psi_i) = \left(\frac{1}{1/\beta + 1/\psi_i}\right)\frac{(\theta^*_{I^*} - \theta^*_i)^2}{2}.$$

To understand this formula, imagine we use a deterministic allocation rule that collects $n\beta$ and $n\psi_i$ observations from I^* and i. Let X_{I^*} and X_i denote the respective sample means. The empirical difference is normally distributed: $X_{I^*} - X_i \sim \mathcal{N}(\Delta, \sigma^2/n)$, where $\Delta = \theta^*_{I^*} - \theta^*_i$ and $\sigma^2 = 1/\beta + 1/\psi_i$. Standard Gaussian tail bounds imply that as $n \to \infty$, $\mathbb{P}(X_{I^*} - X_i < 0) \doteq \exp(-n/2(\sigma\Delta)^2)$, and so $C_i(\beta,\psi_i)$ appears to characterize the probability of error.

The next proposition formalizes the derivations in this section and states that the solution to the above maximization problem attains the optimal error exponent. Recall that $\psi_{n,i} \triangleq \mathbb{P}(I_n = i|\mathcal{F}_{n-1})$ denotes the measurement effort assigned design i at time n.

Proposition 6. *Let* $\boldsymbol{\psi}^*$ *denote the optimal solution to the maximization problem* (11). *If* $\boldsymbol{\psi}_n = \boldsymbol{\psi}^*$ *for all* n, *then*

$$\Pi_n(\Theta^c_{I^*}) \doteq \exp\{-n\Gamma^*\}.$$

Moreover, under any other adaptive allocation rule,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \Pi_n(\Theta^c_{I^*}) \leq \Gamma^*.$$

This shows that under the fixed allocation rule $\boldsymbol{\psi}^*$ error decays as $e^{-n\Gamma^*}$, and that no faster rate of decay is possible, even under an adaptive allocation.

An Optimal Constrained Allocation. Because the algorithms studied in this paper always allocate β–fraction of their samples to measuring I^* in the long run, they may not exactly attain the optimal error exponent. To make rigorous claims about their performance, consider a modified version of the error exponent (11) given by the constrained max-min problem

$$\Gamma^*_\beta \triangleq \max_{\psi:\psi_{I^*}=\beta} \min_{i\neq I^*} C_i(\beta,\psi_i). \tag{12}$$

This optimization problem yields the optimal allocation subject to a constraint that β–fraction of the samples are spent on I^*. The next subsection will show that TTTS, TTPS, and TTVS attain the error exponent Γ^*_β. The next proposition formalizes that the solution to this optimization problem represents an optimal constrained allocation. In addition, it shows that the solution is the unique feasible allocation, under which $C_i(\beta,\psi_i)$ is equal for all suboptimal designs $i \neq I^*$. To understand this result, consider the case where there are three designs and $\theta^*_1 > \theta^*_2 > \theta^*_3$. If $\psi_2 = \psi_3$, then $C_2(\beta,\psi_2) < C_3(\beta,\psi_3)$, reflecting that it is more difficult to certify that $\theta^*_2 \leq \theta^*_{I^*}$ than $\theta^*_3 \leq \theta^*_{I^*}$. The next proposition shows that it is optimal to decrease ψ_2 and increase ψ_1, until the point when $C_2(\beta,\psi_2) = C_3(\beta,\psi_3)$. Instead of allocating *equal measurement effort* to each alternative, it is optimal to adjust measurement effort to gather *equal evidence* to rule out each suboptimal alternative. The results in this proposition are very closely related to those in Glynn and Juneja (2004), in which large deviations rate functions take the place of the functions C_i.

Proposition 7. *The solution to the optimization problem* (12) *is the unique allocation* $\boldsymbol{\psi}^*$ *satisfying* $\psi^*_{I^*} = \beta$ *and*

$$C_i(\beta,\psi^*_i) = C_j\left(\beta,\psi^*_j\right) \qquad \forall\, i,j \neq I^*.$$

If $\boldsymbol{\psi}_n = \boldsymbol{\psi}^*$ *for all* n, *then*

$$\Pi_n(\Theta^c_{I^*}) \doteq \exp\left\{-n\Gamma^*_\beta\right\}.$$

Moreover, under any other adaptive allocation rule, if $\overline{\psi}_{n,I^*} \to \beta$, *then*

$$\limsup_{n\to\infty} -\frac{1}{n}\log \Pi_n(\Theta^c_{I^*}) \leq \Gamma^*_\beta,$$

almost surely.

The following lemma relates the constrained exponent Γ^*_β to Γ^*. This result implies that $e^{-2n\Gamma^*_{1/2}} \leq e^{-n\Gamma^*}$. Therefore, $2n$ samples collected from an algorithm that attains the exponent $\Gamma^*_{1/2}$ give a faster asymptotic rate of posterior convergence than n samples collected using any adaptive algorithm.

Lemma 3. *For* $\beta^* = \arg\max_\beta \Gamma^*_\beta$ *and any* $\beta \in (0,1)$,

$$\frac{\Gamma^*}{\Gamma^*_\beta} \leq \max\left\{\frac{\beta^*}{\beta}, \frac{1-\beta^*}{1-\beta}\right\}.$$

Therefore, $\Gamma^* \leq 2\Gamma^*_{1/2}$.

6.5. Convergence of Top-Two Algorithms

Instead of attempting to directly solve the optimization problem (11), this paper focuses on simple and intuitive sequential strategies. These algorithms have the potential to explore much more intelligently in early stages, as they carefully measure and reason about uncertainty. Although they ostensibly have no connection to the derivations earlier in this section, we establish that, remarkably, all three automatically converge to the unknown optimal allocation. This is shown formally in the next result.

We are now ready to establish the paper's main claim, which shows that TTTS, TTPS, and TTVS each attain the error exponent Γ^*_β.

Proposition 8. *Under the TTTS, TTPS, or TTVS algorithm with parameter $\beta > 0$, $\overline{\boldsymbol{\psi}}_n \to \boldsymbol{\psi}^\beta$, where $\boldsymbol{\psi}^\beta$ is the unique allocation with $\psi^\beta_{I^*} = \beta$ satisfying*

$$C_i\left(\beta, \psi^\beta_i\right) = C_j\left(\beta, \psi^\beta_j\right) \qquad \forall i, j \neq I^*.$$

Therefore,

$$\Pi_n\left(\Theta^c_{I^*}\right) \doteq e^{-n\Gamma^*_\beta}.$$

To understand this result, imagine that n is very large, and $\overline{\psi}_{n,I^*} \approx \beta$. If the algorithm has allocated too much measurement effort to a suboptimal action i, with $\overline{\psi}_{n,i} > \psi^\beta_i + \delta$ for a constant $\delta > 0$, then it must have allocated too little measurement effort to at least one other suboptimal design $j \neq i$. Because much less evidence has been gathered about j than i, we expect $\alpha_{j,n} >> \alpha_{j,i}$. When this occurs, TTTS, TTPS, and TTVS essentially never sample action i until the average effort $\overline{\psi}_{n,i}$ allocated to design i dips back down toward ψ^β_i. This seems to suggest that the algorithm cannot allocate too much effort to any alternative, but that, in turn, implies that it never allocates too little effort to measuring any alternative.

6.6. Asymptotics of the Value Measure

The proof for top-two value sampling relies on the following lemma, which shows that the posterior value of any suboptimal design is logarithmically equivalent to its probability of being optimal.

Lemma 4. *For any $i \neq I^*$, $V_{n,i} \doteq \alpha_{n,i}$.*

Note that by this lemma,

$$\Pi_n\left(\Theta^c_{I^*}\right) = \sum_{i \neq I^*} \alpha_{n,i} \doteq \sum_{i \neq I^*} V_{n,i},$$

and so all of the asymptotic results in this could be reformulated as statements concerning the value assigned to suboptimal alternatives under the posterior.

The lemma is not so surprising, as $V_{n,i} = \int_{\Theta_i} v_i(\boldsymbol{\theta})\pi_n \cdot (\boldsymbol{\theta})d\boldsymbol{\theta}$ differs from $\alpha_{n,i} = \int_{\Theta_i} \pi_n(\boldsymbol{\theta})d\boldsymbol{\theta}$ only because of the function $v_i(\boldsymbol{\theta})$. The $\pi_n(\boldsymbol{\theta})$ term dominates this integral as $n \to \infty$, because it tends to zero at an exponential rate in n, whereas $v_i(\boldsymbol{\theta})$ is a fixed function of n.

7. Further Simulation Experiments

This section presents further simulation results. The focus is not on competitive benchmarking across the wide array of algorithms that have been proposed by researchers in statistics, operations research, and computer science. Although this could be enormously valuable, carrying out such experiments in a fair manner has proved challenging, as these algorithms are often derived under differing modeling assumptions and differing problem objectives, as well as with numerous tuning parameters that muddle comparisons. We instead aim here to focus on gaining clear insight into two questions. Specifically:

1. How robust is the performance of the proposed top-two sampling algorithm to the choice of tuning parameter? Precisely, across a range of problem instances, how does top-two sampling with the default choice of $\beta = 1/2$ compare relative to an omniscient version of the algorithm, which uses the optimal tuning parameter β^* for that instance?
2. How do top-two sampling algorithms, which need to learn and adapt to the long-run optimal sampling proportions on each problem instance $\boldsymbol{\theta}^*$, perform relative to an omniscient policy that knows and tracks the ideal sampling proportions $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$ on each problem instance? The sampling proportions $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$ are those that attain the maximum in Equation (11) defining the optimal exponent Γ^*.

This section presents simulation results across 14 problem settings. To reduce computational burden, as well as simplify the presentation of the results, the section focuses on top-two Thompson sampling and omits the other two variants of top-two sampling. The results reveal strong performance of top-two Thompson sampling with the ad-hoc choice of tuning parameter $\beta = 1/2$. Interestingly, this method also consistently, and often substantially, outperforms the oracle policy $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$.

Each of the 14 experiments investigates a different problem setting as described in Table 1. The problems are divided between those with binary observations and those with standard Gaussian observation noise. For the binary experiments, an independent uniform prior is used, whereas an independent standard normal prior is used for the second experiment. We consider several types of configurations for the arm means. Experiments 10–14 present randomly drawn instances, where each θ^*_i was sampled independently from a standard normal distribution. These were drawn by using the numpy.random.normal function with seeds 1, 2, 3, 4, and 5, respectively. In the configurations labeled "ascending," the arm means

Table 1. Experiment Specifications

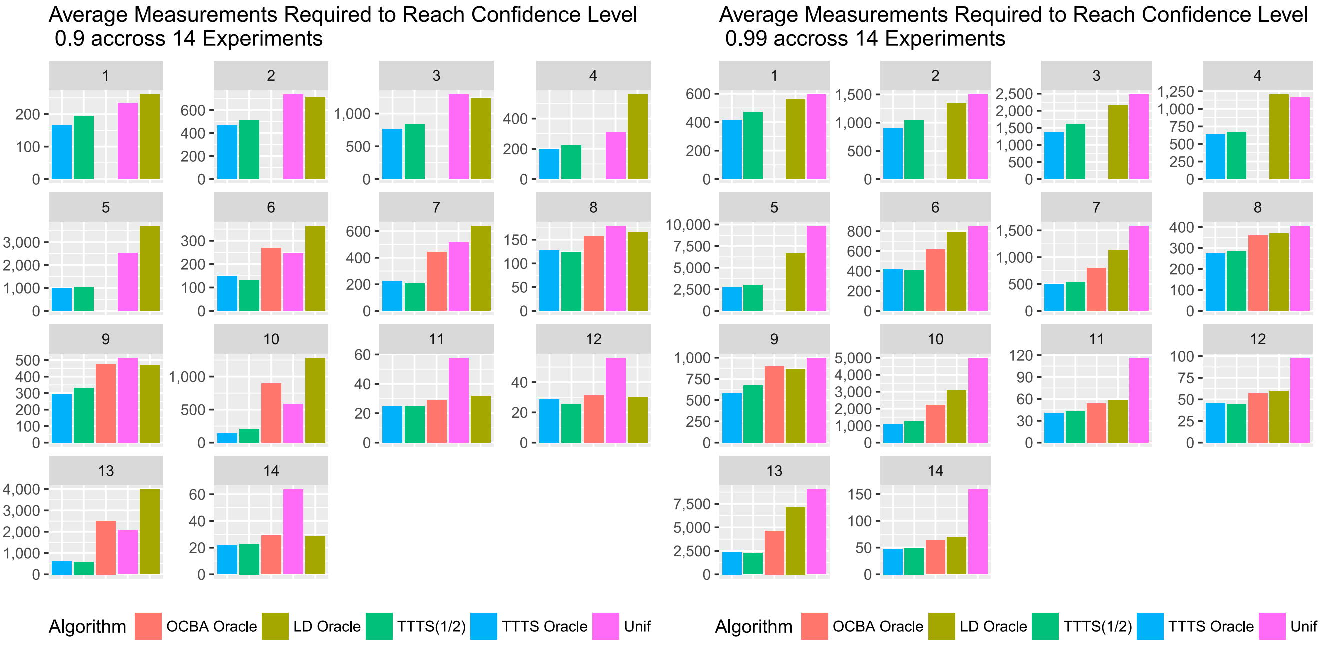
Experiment	Noise	Configuration	k	True arm means $(\theta_1^*, \dots \theta_k^*)$	$\Gamma^*/\Gamma_{\frac{1}{2}}$
1	Binary	Slippage	5	(0.3, 0.3, 0.3, 0.3, 0.5)	1.12
2	Binary	Slippage	10	(0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.5)	1.26
3	Binary	Slippage	15	(0.3, 0.3, 0.3, 0.3, . . . , 0.3, 0.3, 0.3, 0.3, 0.5)	1.34
4	Binary	Ascending	5	(0.1, 0.2, 0.3, 0.4, 0.5)	1.01
5	Binary	Ascending	10	(0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5)	1.01
6	Gaussian	Ascending	5	(−0.5, −0.25, 0, 0.25, 0.5)	1.01
7	Gaussian	Ascending	10	(−0.5, −0.5, −0.5, −0.5, −0.5, −0.5, −0.25, 0, 0.25, 0.5)	1.03
8	Gaussian	Slippage	5	(0, 0, 0, 0, 0.5)	1.11
9	Gaussian	Slippage	10	(0, 0, 0, 0, 0, 0, 0, 0, 0, 0.5)	1.25
10	Gaussian	Random	10	(−2.3, −1.1, −0.8, −0.6, −0.5, −0.2, 0.3, 0.9, 1.6, 1.7)	1.00
11	Gaussian	Random	10	(−2.1, −1.8, −1.2, −1.1, −0.9, −0.8, −0.4, −0.1, 0.5, 1.6)	1.10
12	Gaussian	Random	10	(−1.9, −0.6, −0.5, −0.4, −0.3, −0.1, −0.0, 0.1, 0.4, 1.8)	1.19
13	Gaussian	Random	10	(−1.6, −1.1, −1.0, −0.6, −0.4, 0.1, 0.3, 0.5, 0.6, 0.7)	1.01
14	Gaussian	Random	10	(−0.9, −0.6, −0.3, −0.3, −0.3, 0.1, 0.2, 0.4, 1.6, 2.4)	1.04

increase from lowest to highest with uniform separation between the arms. The slippage configuration was included specifically to investigate cases where top-two sampling performs poorly. In such settings, an equal allocation across arms attains an exponent that is quite competitive, as there are no very poor arms that can be easily ruled out using fewer samples. In addition, the exponent $\Gamma_{\frac{1}{2}}$ attained by TTTS with $\beta = 1/2$ can be farther from the optimal Γ^* than under other problem instances. The ratio of exponents $\Gamma^*/\Gamma_{\frac{1}{2}}$ is displayed for each instance.

Figure 3 displays the average number of measurements required for the posterior to reach a given confidence level. In particular, the experiment tracks the first time when $\max_i \alpha_{n,i} \geq c$ for confidence levels $c = 0.9$ and $c = 0.99$. All results are averaged over 400 trials. This experiment can be thought of as comparing the expected number of samples collected if a natural Bayesian stopping rule is employed. I have chosen to use a Bayesian stopping rule because I do not wish to muddle the comparison between allocation rules by employing a flawed stopping rule that comes with some provable frequentist guarantees. There has been impressive recent progress toward stopping rules that guarantee a frequentist probability of correct selection, but whose stopping regions have a similar shape to the Bayesian one (Garivier and Kaufmann 2016, Kaufmann and Koolen 2018). However, these stopping rules are still highly conservative.

The "large deviations oracle," labeled "LD oracle" in Figure 3, implements the optimal fixed allocation $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$ as prescribed by large deviations theory. At

Figure 3. (Color online) Average Sample Size Required to Reach Confidence Relative to "Oracle" Allocations



each time n, the algorithm constructs the target proportions $n \cdot \psi^*(\boldsymbol{\theta}^*)$ and plays the arm that is most undersampled relative to these proportions. For problems with Gaussian noise, the optimal computing budget allocation of Chen et al. (2000) is a widely used approximation to the fixed allocation $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$. The algorithm labeled OCBA oracle implements the true sampling proportions specified by Chen et al. (2000) for each problem instance. We also compare the uniform or equal allocation, TTTS with tuning parameter $\beta = 1/2$ and TTTS Oracle, which is TTTS with the optimal problem-dependent tuning parameter β^*.

At a high level, there are two key findings from these experiments. In all cases, sample size comparisons refer to the confidence level $c = 0.99$.

1. Top-two Thompson sampling with tuning parameter 1/2 generally offers similar performance to top-two Thompson sampling with the optimal tuning parameter β^*. The most significant separation in performance was on slippage configurations, where TTTS with optimal tuning parameter saved up to 15% of samples on average. On most other instances, using the optimal tuning parameter offered no improvement.
2. The large deviations oracle and the OCBA oracle were consistently, and sometimes dramatically, outperformed. Each one required least 19% more samples on average than TTTS(1/2) for *all 14 experiments.* In their worst experiments, the LD oracle and OCBA oracle used, respectively, more than 200% and 300% the average number of samples used by TTTS(1/2).

The second finding may be quite surprising to some readers. There is a quite a large literature that aims to implement optimal large deviations allocations derived in Glynn and Juneja (2004), or a simpler approximation to these in the Gaussian case known as the OCBA (Chen et al. 2000). Such approaches have also been extended to a number of related problem settings. The allocation $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$ has desirable theoretical properties, including maximizing the asymptotic rate of posterior convergence. A major challenge, however, is that such allocations cannot be directly implemented, as they require knowledge of the true problem instance $\boldsymbol{\theta}^*$. Researchers typically implement an approach that solves for the optimal budget allocation under point estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$, aiming to converge to the prescribed optimal sampling proportions as rapidly as possible. Here, we instead compete against an oracle that knows and carefully follows the asymptotically optimal sampling proportions for each problem instance. Even these oracle policies are significantly outperformed by top-two Thompson sampling with the ad-hoc choice of tuning parameter.

To provide some assurance that this performance gap is not an artifact of the performance criterion, we ran further experiments focused on two alternative performance measures. For each trial and each algorithm, we tracked the identity of the arm with highest empirical mean as measurements were gathered. Figure 4 displays the probability of incorrect selection and expected simple regret (Bubeck et al. 2009), averaged across 1,000 trials. Figure 4(a), Figure 4(b), and Figure 4(c), respectively, correspond to problems 4, 5, and 2 from Table 1. Confidence bands are hardly visible to the naked eye, and hence are omitted. Again, according to these experiments, top-two Thompson sampling substantially outperforms the static allocation $\psi^*(\boldsymbol{\theta}^*)$, to which it converges asymptotically. One stark feature of the experiments is the high "simple regret" incurred under the $\boldsymbol{\psi}^*(\boldsymbol{\theta}^*)$ allocation. This, apparently, is because the allocation sometimes mistakenly identifies highly inferior arms as optimal, whereas when the uniform allocation misidentifies the optimal arm, it still tends to return a near-optimal one.

Current theory does not explain why top-two Thompson sampling appears to outperform the static oracle allocations in these experiments. It is worth offering some possible intuition, however. First, the oracle allocations are based on a number of approximations, either in the form of tail approximations to the posterior of each arm or certain union bounds. By contrast, Thompson sampling uses exact samples from the posterior distribution and may more accurately reflect uncertainty in early stages. Second, even if the oracle allocations know the true-arm means, they do not adapt in response to unusual observations. Thompson sampling, on the other hand, is fully adaptive, and can gather fewer samples from an arm if early samples provide strong evidence that arm is suboptimal.

8. Extensions and Open Problems

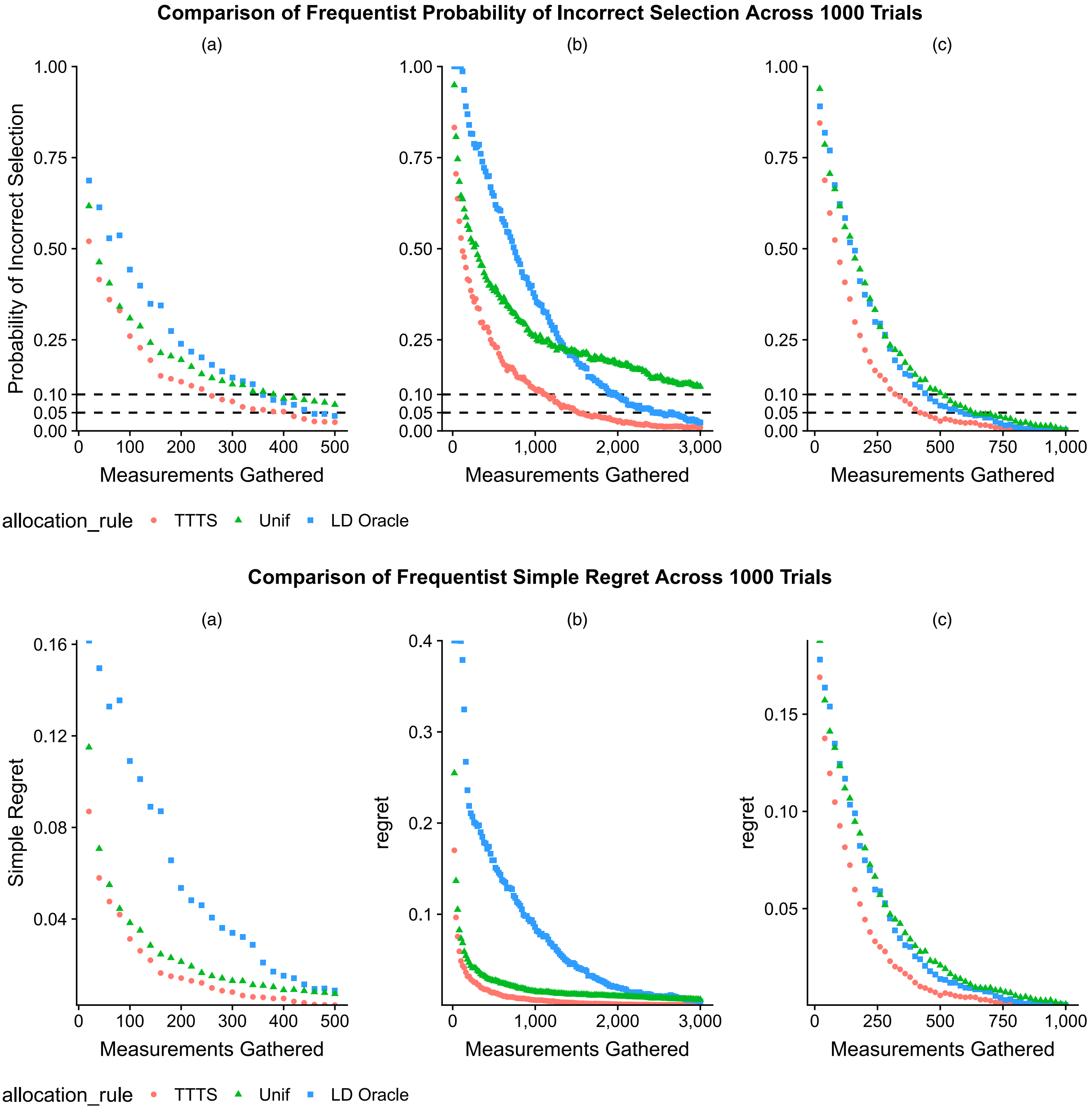
This paper studies efficient adaptive allocation of measurement effort for identifying the best among a finite set of options or designs. We propose three simple Bayesian algorithms. Each is a variant of what we call top-two sampling, which, at each time step, measures one of the two designs that appear most promising given current evidence. Surprisingly, these seemingly naive algorithms are shown to satisfy a strong asymptotic optimality property.

Top-two sampling appears to be a general design principle that can be extended to address a variety of problems beyond to the scope of this paper. To spur research in this area, we briefly discuss a number of extensions and open questions below.

8.1. Top-Two Sampling Via Constrained Maximum a Posteriori Estimation

Here, we present a version of top-two sampling that uses maximum a posteriori (MAP) estimation. This can simplify computations, as MAP estimates can be computed without solving for the normalizing

Figure 4. (Color online) Comparison of Frequentist Probability of Incorrect Selection and Simple Regret on Three Problem Instances with Binary Observations



Note. Panels (a), (b), and (c), respectively, correspond to Experiments 4, 5, and 2 from Table 1.

constant of the posterior density $\pi_n(\boldsymbol{\theta})$. Consider the following procedure for selecting a design at time n:

1. Compute $\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\Theta} \pi_n(\boldsymbol{\theta})$ and set $\hat{I}_n = \arg\max_i \hat{\theta}_i$.

2. Compute $\hat{\boldsymbol{\theta}}' \in \arg\max_{\boldsymbol{\theta}\in\Theta^c_{\hat{I}_n}} \pi_n(\boldsymbol{\theta})$ and set $\hat{J}_n = \arg\max_i \hat{\theta}'_i$.

3. Play $(\hat{I}_n, \hat{J}_n)$ with respective probabilities $(\beta, 1-\beta)$.

The first step uses MAP estimation to make a prediction $\hat{I}_n$ of the best design, whereas the second uses constrained MAP estimation to identify the alternative design that is most likely to be optimal when $\hat{I}_n$ is not. Many of the asymptotic calculations in the previous section appear to extend to this algorithm, but proving this formally is left as an open problem.

8.2. Indifference-Zone Criterion

Suppose our goal is to confidently identify an ϵ–optimal arm, for a user-specified indifference parameter $\epsilon > 0$.

Much of the paper investigates the set of parameters Θ_i under which arm i is optimal and studies the rate at which $\Pi_n(\Theta_{I^*}) \to 1$. Now, let us instead consider the set of parameters

$$\Theta_{\epsilon,i} = \left\{\boldsymbol{\theta} | \theta_i \ge \max_j \theta_j - \epsilon\right\},$$

under which i is ϵ–optimal. It is easy to develop a variety of modified top-two sampling rules under which $\max_i \Pi_n(\Theta_{\epsilon,i}) \to 1$ rapidly. For example, we can extend TTPS as follows: Set $\hat{I}_n = \arg\max_i \Pi_n(\Theta_{\epsilon,i})$. Define $\hat{J}_n = \arg\max_{j \neq \hat{I}_n} \Pi_n(\boldsymbol{\theta} | \theta_j = \max_i \theta_i \,\&\, \theta_j > \theta_{\hat{I}_n} + \epsilon)$ to be the alternative design that is most likely to be optimal and offer an ϵ-improvement over $\hat{I}_n$. A top-two Thompson sampling approach might instead continue sampling $\boldsymbol{\theta} \sim \Pi_n$ until $\max_i \theta_i > \theta_{\hat{I}_n} + \epsilon$ and then set $J_n = \arg\max_i \theta_i$.

8.3. Top *m*-Arm Identification

Suppose now that our goal is to identify the top $m < k$ designs. Consider choosing a design to measure at time n by the following steps:

1. Sample $\boldsymbol{\theta} \sim \Pi_n$ and compute the top m designs under $\boldsymbol{\theta}$.
2. Continue sampling $\boldsymbol{\theta}' \sim \Pi_n$ until the top m designs under $\boldsymbol{\theta}'$ differ from those under $\boldsymbol{\theta}$.
3. Identify the set of designs that are in the top m under $\boldsymbol{\theta}$ or under $\boldsymbol{\theta}'$, but not under both. Choose a design to measure by sampling one uniformly at random from this set.

This is the natural extension of top-two Thompson sampling to the top-m arm problem. In fact, when $m = 1$, this is exactly TTTS with $\beta = 1/2$. I conjecture that, like the case where $m = 1$, this algorithm attains a rate of posterior convergence within a factor of 2 of optimal for general m. The optimal exponent for this problem can be calculated by mirroring the steps in Section 6.4.

8.4. Extremely Correlated Designs

Although our results apply in the case of correlated priors, the proposed algorithms may be wasteful when there are a large number of designs whose qualities are extremely correlated. As an example, consider an extension of our techniques to a pure-exploration variant of a linear bandit problem. Here, we associate each action i with a feature vector $x_i \in \mathbb{R}^d$ and seek an action that maximizes $x_i^T\theta$. The vector $\theta \in \mathbb{R}^d$ is unknown, but we begin with a prior $\theta \sim N(0, I)$ and see noisy observations of $x_i^T\theta$ whenever action i is selected. To apply top-two sampling to this problem, we should modify the algorithm's second step. For example, under top-two Thompson sampling, we usually begin drawing a design according to $\hat{i} \sim \boldsymbol{\alpha}_n$, and then continue drawing designs $\hat{j} \sim \boldsymbol{\alpha}_n$ until $\hat{i} \neq \hat{j}$. These are played with respective probabilities $(\beta, 1-\beta)$. But even if $\hat{i} \neq \hat{j}$, their features may be nearly identical. A more natural extension of top-two Thompson sampling would modify the second step, and continue sampling $\hat{j} \sim \boldsymbol{\alpha}_n$, until a sufficiently different action is drawn—for example, until the angle between $x_{\hat{j}}$ and $x_{\hat{i}}$ exceeds a threshold.

8.6. Tuning β

The most glaring gap in this work may be arbitrary choice of tuning parameter β. Optimal asymptotic rates can be attained by adjusting this parameter over time by solving for an optimal allocation, as in (11). It is an open problem to instead develop simple algorithms that set β automatically through value of information calculations, or avoid the need for such a parameter altogether.

8.7. Adaptive Stopping

This paper proposed only an allocation rule, which determines the sequence of measurements to draw, but this can be coupled with a rule that determines when to stop sampling. One natural stopping rule in a Bayesian framework is to stop when $\max_i \alpha_{n,i} > 1 - \delta$ for some $\delta > 0$. Let τ_δ be a random variable indicating the stopping time under constraint δ. Because $1 - \max_i \alpha_{n,i} \doteq e^{-n\Gamma^*_\beta}$ under top-two sampling, our results imply that for each sample path $\tau_\delta \sim \Gamma^*_\beta \log(1/\delta)$ as $\delta \to 0$. It is natural to conjecture that $\mathbb{E}[\tau_\delta] \sim \Gamma^*_\beta \log(1/\delta)$ as well. This closely mirrors optimal results in Chernoff (1959), Jennison et al. (1982), and Kaufmann (2018). Does this rule also yield a frequentist probability of incorrect selection that is $O(\delta)$ as $\delta \to 0$? More generally, an open problem is to show that, when combined with an appropriate stopping rule, top-two sampling schemes nearly minimize the expected number of samples $\mathbb{E}[\tau_\delta]$, as in Jennison et al. (1982) or Kaufmann (2018). A follow-up to the current paper has addressed this for a particular top-two sampling algorithm in the case of Gaussian observations (Qin et al. 2017).

Acknowledgments

The author thanks Shane Henderson, Chao Qin, and Kevin Jamieson for careful reading and many thoughtful comments on the paper. Sandeep Juneja pointed out to the author how the sample complexity terms differ in the fixed-budget and fixed-confidence settings, which was previously misstated in the literature review. The author also thanks anonymous reviewers for *Operations Research* and COLT for many comments that helped improve the paper's presentation, as well as many seminar participants.

Endnotes

[1] Interpreted in the context of clinical trials, this paper's results are stated in terms of the number of patients required to reach a confident conclusion of the best treatment. However, we will see that optimal rules from this perspective also allocate fewer patients to very poor treatments, potentially leading to more ethical trials (Berry 2004).

[2] TTVS is executed with the utility function $u(\theta) = \theta$.

[3] See, for example, Qin et al. (2017), which is a follow-up to the current paper.

[4] For Gaussian distributions, this is exactly the exponent presented in Glynn and Juneja (2004). In general, it differs. Our exponent depends on Kullback–Leibler (KL) divergences of the form $d(\theta_i^*, \theta_i)$, which is mirrors the optimal sample complexity terms in the fixed confidence setting (Chernoff 1959, Jennison et al. 1982, Glynn and Juneja 2004). The exponent in Glynn and Juneja (2004) is derived for the fixed budget setting and depends everywhere on flipped KL divergence terms of the form $d(\theta_i \| \theta_i^*)$.

References

Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. Mannor S, Srebro N, Williamson RC, eds. *Proc. 21st Annual Conf. Learning Theory*, Proceedings of Machine Learning Research, vol. 23 (PMLR), 39.1–39.26.

Albert AE (1961) The sequential design of experiments for infinitely many states of nature. *Ann. Math. Statist.* 32(3):774–799.

Audibert JY, Bubeck S, Munos R (2010) Best arm identification in multi-armed bandits. Kalai AT, Mohri M, eds. *COLT 23rd Conf. Learning Theory* (Omnipress, Madison, WI), 41–53.

Barron A, Schervish MJ, Wasserman L (1999) The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* 27(2):536–561.

Bechhofer RE, Sobel M(1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.* 25(2):16–39.

Berry DA (2004) Bayesian statistics and the efficiency and ethics of clinical trials. *Statist. Sci.* 19(1):175–187.

Bubeck S, Munos R, Stoltz G (2009) Pure exploration in multi-armed bandits problems. *Algorithmic Learning Theory*, Lecture Notes in Computer Science, vol. 5809 (Springer, Berlin), 23–37.

Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, eds. *Proc. Adv. Neural Inform. Processing Systems NIPS 2011* (Curran Associates, Red Hook, NY), 2249–2257.

Chen CH, Chick SE, Lee LH, Pujowidianto NA (2015) Ranking and selection: Efficient simulation budget allocation. Fu MC, ed. *Handbook of Simulation Optimization*, International Series in Operations Research & Management Science, vol. 216 (Springer, New York), 45–80.

Chen CH, He D, Fu M, Lee LH (2008) Efficient simulation budget allocation for selecting an optimal subset. *INFORMS J. Comput.* 20(4):579–595.

Chen CH, Lin J, Yücesan E, Chick SE (2000) Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynam. Systems* 10(3):251–270.

Chernoff H (1959) Sequential design of experiments. *Ann. Math. Statist.* 30(3):755–770.

Chernoff H (1975) Approaches in sequential design of experiments. Srivastava JN, ed. *A Survey of Statistical Design and Linear Models* (North-Holland, Amsterdam), 67–90.

Chick SE, Frazier P (2012) Sequential sampling with economics of selection procedures. *Management Sci.* 58(3):550–569.

Chick SE, Gans N (2009) Economic analysis of simulation selection problems. *Management Sci.* 55(3):421–437.

Chick SE, Inoue K (2001) New two-stage and sequential procedures for selecting the best simulated system. *Oper. Res.* 49(5):732–743.

Chick SE, Branke J, Schmidt C (2010) Sequential sampling to myopically maximize the expected value of information. *INFORMS J. Comput.* 22(1):71–80.

Diaconis P, Freedman D (1986) On the consistency of Bayes estimates. *Ann. Statist.* 14(1):1–26.

Even-Dar E, Mannor S, Mansour Y (2002) PAC bounds for multi-armed bandit and Markov decision processes. Kivinen J, Sloan RH, eds. *COLT '02 Proc. 15th Annual Conf. Comput. Learn. Theory*, Lecture Notes in Computer Science, vol. 2375 (Springer, Berlin), 255–270.

Fan W, Hong LJ, Nelson BL (2016) Indifference-zone-free selection of the best. *Oper. Res.* 64(6):1499–1514.

Ferreira KJ, Simchi-Levi D, Wang H (2018) Online network revenue management using Thompson sampling. *Oper. Res.* 66(6): 1586–1602.

Frazier PI (2014) A fully sequential elimination procedure for indifference-zone ranking and selection with tight bounds on probability of correct selection. *Oper. Res.* 62(4):926–942.

Frazier PI, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM J. Control Optim.* 47(5):2410–2439.

Freedman DA (1963) On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.* 34(4):1386–1403.

Gabillon V, Ghavamzadeh M, Lazaric A (2012) Best arm identification: A unified approach to fixed budget and fixed confidence. Pereira F, Burges CJC, Bottou L, Weinberger KQ, eds. *NIPS'12 Proc. 25th Internat. Conf. Neural Inform. Processing Systems*, vol. 2 (Curran Associates, Red Hook, NY), 3212–3220.

Garivier A, Kaufmann E (2016) Optimal best arm identification with fixed confidence. Feldman V, Rakhlin A, Shamir O, eds. *Proc. Conf. Learn. Theory (COLT) 2016*, Proceedings of Machine Learning Research, vol. 49 (PMLR), 1028–1050.

Ghosal S, Ghosh JK, Van Der Vaart AW (2000) Convergence rates of posterior distributions. *Ann. Statist.* 28(2):500–531.

Gittins JC (1979) Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. B.* 41(2):148–164.

Gittins JC, Jones DM (1974) A dynamic allocation index for the sequential design of experiments. Gani J, ed., *Progress in Statistics* (North-Holland, Amsterdam), 241–266.

Glynn P, Juneja S (2004) A large deviations perspective on ordinal optimization. Ingalls RG, Rossetti MD, Smith JS, Peters BA, eds. *Proc. 2004 Winter Simulation Conf. 2004*, vol. 1 (IEEE, Piscataway, NJ).

Glynn P, Juneja S (2015) Selecting the best system and multi-armed bandits. Preprint, submitted July 16, https://arxiv.org/abs/1507.04564.

Gopalan A, Mannor S, Mansour Y (2014) Thompson sampling for complex online problems. Xing EP, Jebara T, eds. *ICML 14 Proc. 31st Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 32 (PMLR), 100–108.

Graepel T, Candela JQ, Borchert T, Herbrich R (2010) Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. Fürnkranz J, Joachims T, eds. *Proc. 27th Internat. Conf. Machine Learn. (ICML-10)* (Omnipress, Madison, WI), 13–20.

Gupta SS, Miescke KJ (1996) Bayesian look ahead one-stage sampling allocations for selection of the best population. *J. Statist. Planning Inference* 54(2):229–244.

Gutin E, Farias VF (2016) Optimistic Gittins indices. Lee DD, von Luxburg U, Garnett R, Sugiyama M, Guyon I, eds. *NIPS'16 Proc. 30th Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 3161–3169.

Hong LJ, Nelson BL, Xu J (2015) Discrete optimization via simulation. Fu MC, ed. *Handbook of Simulation Optimization*, International Series in Operations Research and Management Science, vol. 216 (Springer, New York), 9–44.

Hunter S, Pasupathy R (2013) Optimal sampling laws for stochastically constrained simulation optimization on finite sets. *INFORMS J. Comput.* 25(3):527–542.

Jamieson K, Nowak R (2014) Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. *48th Annual Conf. Inform. Sci. Systems (CISS)* (IEEE, Piscataway, NJ), 1–6.

Jennison C, Johnstone IM, Turnbull BW (1982) Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. Gupta SS, Berger JO, eds. *Statistical Decision Theory and Related Topics III* (Academic Press, New York), 55–86.

Karnin Z, Koren T, Somekh O (2013) Almost optimal exploration in multi-armed bandits. Dasgupta S, McAllester D, eds. *Proc. 30th Internat. Conf. Machine Learn.*, Proceedings of Machine Learning Research, vol. 28 (PMLR), 1238–1246.

Kaufmann E (2018) On Bayesian index policies for sequential resource allocation. *Ann. Statist.* 46(2):842–865.

Kaufmann E, Kalyanakrishnan S (2013) Information complexity in bandit subset selection. *J. Machine Learn. Res*. 30:228–251.

Kaufmann E, Koolen W (2018) Mixture Martingales revisited with applications to sequential tests and confidence intervals. Preprint, submitted November 28, https://arxiv.org/abs/1811.11419.

Kaufmann E, Cappé O, Garivier A (2012) On Bayesian upper confidence bounds for bandit problems. Lawrence ND, Girolami MA, eds. *Proc. 15th Internat. Conf. Artificial Intelligence Statist.*, Proceedings of Machine Learning Research, vol. 22 (PMLR), 592–600.

Kaufmann E, Cappé O, Garivier A (2014) On the complexity of best-arm identification in multi-armed bandit models. *J. Machine Learning Res*. 17.1(2016):1–42.

Kaufmann E, Korda N, Munos R (2012) Thompson sampling: An asymptotically optimal finite time analysis. Bshouty NH, Stoltz G, Vayatis N, Zeugmann T, eds. *Internat. Conf. Algorithmic Learn. Theory*, Lecture Notes in Computer Science, vol. 7568 (Springer, Berlin), 199–213.

Keener R (1984) Second order efficiency in the sequential design of experiments. *Ann. Statist.* 12(2):510–532.

Kiefer J, Sacks J (1963) Asymptotically optimum sequential inference and design. *Ann. Math. Statist.* 34(3):705–750.

Kim S-H, Nelson BL (2006) Selecting the best system. Henderson SG, Nelson BL, eds. *Simulation*, Handbooks in Operations Research and Management Science, vol. 13 (Elsevier, Amsterdam), 501–534.

Kim S-H, Nelson BL (2007) Recent advances in ranking and selection. Henderson S, Biller B, Hsieh M-h, Shortle J, eds. *Proc. 39th Conf. Winter Simulation: 40 Years! Best Is Yet to Come* (IEEE, Piscataway, NJ), 162–172.

Korda N, Kaufmann E, Munos R (2013) Thompson sampling for one-dimensional exponential family bandits. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, eds. *NIPS'13 Proc. 26th Internat. Conf. Neural Inform. Processing Systems*, vol. 1 (Curran Associates, Red Hook, NY), 1448–1456.

Lai T, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22.

Mannor S, Tsitsiklis JN (2004) The sample complexity of exploration in the multi-armed bandit problem. *J. Machine Learn. Res.* 5: 623–648.

Naghshvar M, Javidi T (2013) Active sequential hypothesis testing. *Ann. Statist.* 41(6):2703–2738.

Ni EC, Ciocan DF, Henderson SG, Hunter SR (2017) Efficient ranking and selection in parallel computing environments. *Oper. Res.* 65(3):821–836.

Nitinawarat S, Atia GK, Veeravalli VV (2013) Controlled sensing for multihypothesis testing. *IEEE Trans. Automat. Control*. 58(10): 2451–2464.

Pasupathy R, Hunter SR, Pujowidianto NA, Lee LH, Chen C (2015) Stochastically constrained ranking and selection via score. *ACM Trans. Model. Comput. Simul. (TOMACS)* 25(1):Article 1.

Paulson E (1964) A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Statist.* 35(1):174–180.

Qin C, Klabjan D, Russo D (2017) Improving the expected improvement algorithm. von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, eds. *NIPS'17 Proc. 31st Internat. Conf. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 5387–5397.

Rinott Y (1978) On two-stage selection procedures and related probability-inequalities. *Comm. Statist. Theory Methods* 7(8): 799–811.

Russo D, Van Roy B (2017) Learning to optimize via information-directed sampling. *Oper. Res.* 66(1):230–252.

Ryzhov IO (2016) On the convergence rates of expected improvement methods. *Oper. Res.* 64(6):1515–1528.

Ryzhov IO, Powell WB, Frazier PI (2012) The knowledge gradient algorithm for a general class of online learning problems. *Oper. Res.* 60(1):180–195.

Scott SL (2016) Overview of content experiments: Multi-armed bandit experiments. Accessed November 9, 2016, https://support.google.com/analytics/answer/2844870?hl=en.

Srinivas N, Krause A, Kakade S, Seeger M (2012) Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Trans. Inform. Theory* 58(5): 3250–3265.

Tang L, Rosales R, Singh A, Agarwal D (2013) Automatic ad format selection via contextual bandits. *Proc. 22nd ACM Internat. Conf. Inform. Knowledge Management* (ACM, New York), 1587–1594.

Thompson W (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

Villar SS, Bowden J, Wason J (2015) Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statist. Sci.* 30(2):199–215.

Daniel Russo is an assistant professor in the Decision, Risk, and Operations Division of Columbia Business School. His research lies at the intersection of statistical machine learning and sequential decision making and contributes to the fields of online optimization and reinforcement learning.