

# Simple Bayesian Algorithms for Best-Arm Identification

Daniel Russo

December 9, 2016

## 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 which considers not only the probability 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 the posterior definitively identifies the optimal design, in the sense 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.

## 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 multi-armed 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, and 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.<sup>1</sup> Multi-armed 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 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 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 multi-armed bandits. Thompson sampling has attracted a great deal of recent interest in both academia and industry [[Scott, 2016](#), [Tang et al., 2013](#), [Graepel et al., 2010](#), [Chapelle and Li, 2011](#), [Agrawal and Goyal, 2012](#), [Kauffmann et al., 2012](#), [Gopalan et al., 2014](#), [Russo and Van Roy, 2014](#)], 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, 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, and the corresponding exponent is essentially the best

---

<sup>1</sup>Interpreted 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](#)].

possible among all allocation rules.

## 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 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 7 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 a sharp result of this type by proving that three top-two sampling rules attain an optimal rate of posterior convergence across a broad class of problems. In establishing this result, 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 non-conjugate priors and under adaptive measurement rules.

## 1.2 Related Literature

**Sequential Bayesian Best-Arm Identification.** There is a sophisticated literature on algorithms for Bayesian multi-armed 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, 2016], Thompson sampling [Agrawal and Goyal, 2012, Korda et al., 2013, Gopalan et al., 2014, Johnson et al., 2015], information-directed sampling [Russo and Van Roy, 2014], 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 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. See appendix C 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*. See Kim and Nelson [2006] and Kim and Nelson [2007] for 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 [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 a design within  $\epsilon$  of optimal. Assuming measurement noise is Gaussian with known variance  $\sigma^2$ , one can guarantee this indifference-zone criterion by gathering  $O((\sigma k/\epsilon^2) \log(k/\delta))$  total measurements, divided equally among the  $k$  designs, and then returning the design with 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] shows that when measurement noise is uniformly bounded, the indifference-zone criterion is satisfied by a sequential elimination strategy that uses only  $O((k/\epsilon^2) \log(1/\delta))$  samples on average. Mannor and Tsitsiklis [2004] provide a matching lower bound. Similar to minimax bounds, this shows the upper bound of Even-Dar et al. [2002] is tight, up to a constant factor, for a certain worst case problem instance. 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 7 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 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 – i.e. 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 optimal-computing-budget allocation (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 multi-armed bandit literature. Unfortunately, it may be difficult to implement the allocation in Glynn and Juneja [2004] without additional prior knowledge. Later work by Glynn and Juneja [2015] provides a substantial 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 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]. Rather than take 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, Nitinawarat et al., 2013, Naghshvar 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  $\delta \rightarrow 0$ . 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 multi-armed 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 and Bubeck, 2010, Gabillon et al., 2012, Karnin et al., 2013, Jamieson and Nowak, 2014]. However, upper and lower bounds match only up 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]. This exponent is then simplified into a form that mirrors one derived by Glynn and Juneja [2004], and, for Gaussian distributions, one derived by Jennison et al. [1982]. While 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 certain rates are attainable asymptotically, and as Chernoff [1975] writes, “sidestep the issue of how to experiment in the early stages.” By contrast, we show simple and natural adaptive rules automatically reach a notion of asymptotically optimal performance. See Subsection 6.5 for a more precise discussion of these approaches.

## 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  $I_n \in \{1, \dots, k\}$ , 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  $\mathbf{Y}_n \triangleq (Y_{n,1}, \dots, Y_{n,k})$  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  $I_1, Y_{1,I_1}, \dots, I_{n-1}, Y_{n-1,I_{n-1}}$ . 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  $\nu$ , where  $\theta_i^* \in \mathbb{R}$  is an unknown parameter associated with design  $i$ . This density takes the form

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

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  $\mu(\theta) \triangleq \int yp(y|\theta)d\nu(y)$  is a strictly increasing function of  $\theta$ . 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,  $\boldsymbol{\theta}^* \triangleq (\theta_1^*, \dots, \theta_k^*)$  will denote the unknown true parameter vector, and  $\boldsymbol{\theta}$  and  $\boldsymbol{\theta}'$  will be used to denote possible alternative parameter vectors. Let  $I^* = \arg \max_{1 \leq i \leq k} \theta_i^*$  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  $\epsilon$ -optimal design, for some specified tolerance level  $\epsilon > 0$ .

**Prior and Posterior Distributions.** The policies studied in this paper make use of a prior distribution  $\Pi_1$  over a set of possible parameters  $\Theta$  that contains  $\boldsymbol{\theta}^*$ . Based on a sequence of observations  $(I_1, Y_{1,I_1}, \dots, I_{n-1}, Y_{n-1,I_{n-1}})$ , beliefs are updated to attain a posterior distribution  $\Pi_n$ . We assume  $\Pi_1$  has density  $\pi_1$  with respect to Lebesgue measure. In this case, the posterior distribution  $\Pi_n$  has corresponding density

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

where

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

is the likelihood function. While 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.

**Optimal Action Probabilities.** Let

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

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

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

denote the posterior probability assigned to the event that action  $i$  is optimal. Our analysis will focus on  $\Pi_n(\Theta_{I^*}^c) = 1 - \alpha_{I^*}$ , 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  $\Pi_n(\Theta_{I^*}^c) \rightarrow 0$  as  $n \rightarrow \infty$ , and the rate of convergence is essentially optimal.

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

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

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  $\hat{I} \in \{1, \dots, k\}$  of the optimal design, and measures that with probability  $\beta$ . 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  $\beta$  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 (TTPS)

With probability  $\beta$ , 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 (TTVS)

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 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  $\beta$  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  $\theta_i$ . 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  $i \in \{1, \dots, k\}$ , the function

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

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

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

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  $n \rightarrow \infty$ , 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 multi-armed 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  $\alpha_{n,i} = .95$ , then the algorithm will only select an action other than  $i$  roughly once every 20 periods, greatly extending the time it takes until  $\alpha_{n,i} > .99$ . 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 multi-armed 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 (TTTS)

This section proposes top-two Thompson sampling (TTTS), which modifies standard Thompson sampling by adding a re-sampling 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  $I \sim \alpha_n$ . Design  $I$  is measured with probability  $\beta$ , 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  $J \sim \alpha_n$  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  $\alpha_{n,I^*} \rightarrow 1$ ,  $\psi_{n,I^*} \rightarrow \beta$ , 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  $\beta$ . 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  $\Pi_n$ , it is easy to sample from the posterior distribution over the optimal design  $\alpha_n$ . In particular, if  $\hat{\theta} \sim \Pi_n$  is drawn randomly from the posterior, then  $\arg \max_i \hat{\theta}_i$  is a random sample from  $\alpha_n$ . 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. It is worth highlighting that this algorithm does not require computing or approximating the distribution  $\alpha_n$ .

---

#### Algorithm 1 Top-Two Thompson Sampling ( $\beta$ )

---

- 1: Sample  $\hat{\theta} \sim \Pi_n$  and set  $I \leftarrow \arg \max_i \hat{\theta}_i$  ▷ Apply Thompson sampling
  - 2: Sample  $B \sim \text{Bernoulli}(\beta)$
  - 3: **if**  $B = 1$  **then** ▷ Occurs with probability  $\beta$ .
  - 4:     Play  $I$
  - 5: **else**
  - 6:     **repeat**
  - 7:         Sample  $\hat{\theta} \sim \Pi_n$  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  $\Pi_n$  has a closed form. Algorithm 2 shows how to approximate  $\alpha_{n,i}$  and  $V_{n,i}$  using samples  $\theta^1 \dots \theta^M$ , which enables efficient approximations to TTPS and TTVS whenever posterior samples can be efficiently generated.

---

**Algorithm 2** SampleApprox( $K, M, u, \theta^1, \dots, \theta^M$ )

---

- 1:  $\mathcal{S}_i \leftarrow \{m | i = \arg \max_j \theta_j^m\} \quad \forall i \in \{1, \dots, K\}$
  - 2:  $\hat{\alpha}_i \leftarrow |\mathcal{S}_i|/m \quad \forall i \in \{1, \dots, K\}$
  - 3:  $\hat{V}_i \leftarrow M^{-1} \sum_{m \in \mathcal{S}_i} \left( u(\theta_i^m) - \max_{j \neq i} u(\theta_j^m) \right) \quad \forall i \in \{1, \dots, K\}$
  - 4: **return**  $\hat{\alpha}, \hat{V}$
- 

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

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

where  $f_1$  denotes the PDF of  $X_1$  and  $F_2, \dots, F_K$  are the CDFs of  $X_2, \dots, X_k$ . In particular,  $\mathbf{P}(X_1 = \max_i X_i)$  can be computed by solving a 1-dimensional integral. Based on this insight, Appendix B provides an efficient implementation of TTPS and TTVS for a problem with independent Beta priors and binary observations. That implementation approximates integrals like (5) using quadrature with  $n$  points, and has the time and space complexity that scale as  $O(kn)$ .

## 4 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  $\theta^* = (.1, .2, .3, .4, .5)$  defines the true success probability of each design. Each algorithm begins with an independent uniform prior over the components of  $\theta^*$ . The experiment compares the performance of top-two probability sampling (TTPS), top-two value sampling (TTVS)<sup>2</sup>, and top-two Thompson sampling (TTTS) 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 an especially natural benchmark, as it is the most commonly used strategy in practice.

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 using fewer than half the measurements required by a uniform allocation rule. While 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. Panel (a) compares Thompson sampling to Top-Two Thompson sampling. TS appears to reach low confidence levels as rapidly as top-two TS, but as suggested in Subsection 3.3, is very slow to reach high levels of confidence. It requires over than 60% more measurements to reach

---

<sup>2</sup>TTVS is executed with the utility function  $u(\theta) = \theta$

confidence .95 and over 250% more measurements to reach confidence .99. TS requires an onerous number of measurements to reach confidence .999, and so we omit this experiment.

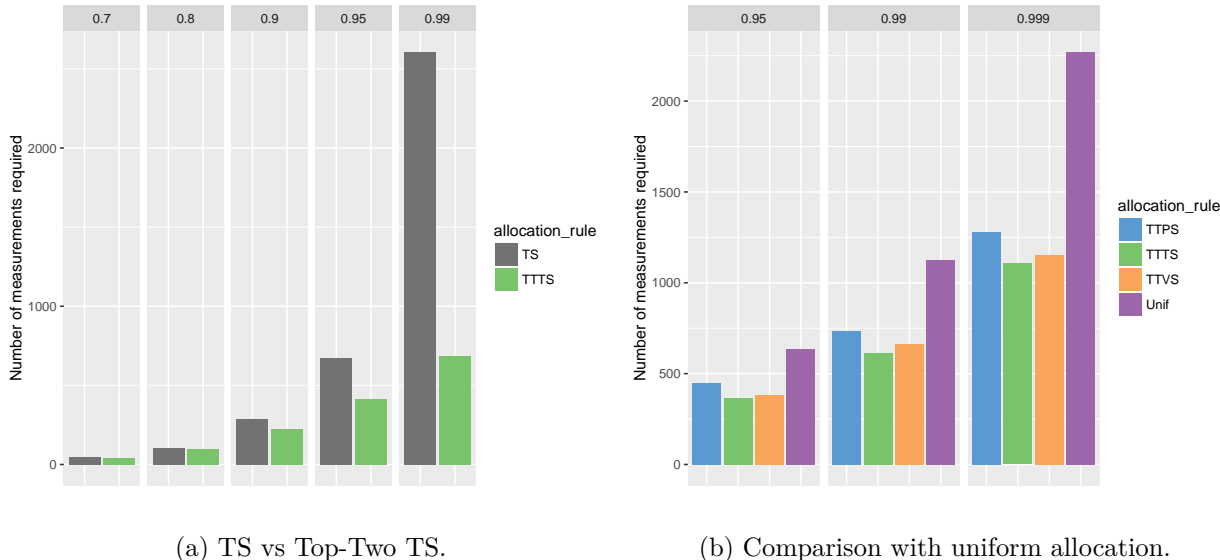
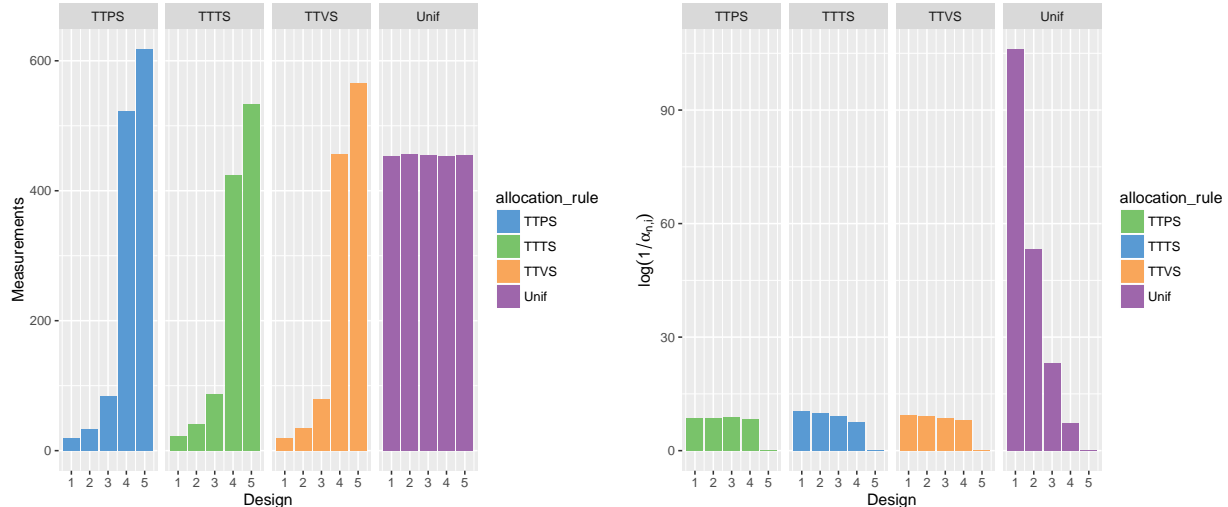


Figure 1: Number of measurements required to reach given confidence level.

Figure 2 provides insight into how the proposed algorithms differ from the uniform allocation. It displays the distribution of measurements and posterior beliefs at the first time when a confidence level of .999 is reached. Again, all results are averaged across 500 trials. Panel (a) displays the average number of measurements collected from each design. It is striking that although TTPS, TTVS, and TTTS 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  $\theta_4^*$  is closer to  $\theta_5^*$ , and because observations from a Bernoulli distribution with parameter .4 have higher variance than under a Bernoulli distribution with parameter .1.

Panel (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 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, TTPS, TTVS, and TTTS 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.



(a) Measurements collected of each design.

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

Figure 2: Distribution of measurements and posterior beliefs at termination.

## 5 Main Theoretical Results

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

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

captures the posterior mass assigned to the event that an action other than  $I^*$  is optimal. One hopes that  $\Pi_n(\Theta_{I^*}^c) \rightarrow 0$  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,  $\Pi_n(\Theta_{I^*}^c)$  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 the use of multivariate Gaussian prior. However, we otherwise allow for quite general correlated priors, expressed in terms of a density over a compact set. Assumption 1 is used only in establishing posterior concentration results, and it is likely that these can be established under less restrictive technical conditions.

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

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

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

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 will be explicitly characterized in Section 6.

The first part of the theorem shows that there is an exponent  $\Gamma^* > 0$  such that  $\Pi_n(\Theta_{I^*}^c)$  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  $\beta$  is set optimally.

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  $\Pi_n(\Theta_{I^*}^c) \rightarrow 0$  at rate  $e^{-n\Gamma_\beta^*}$  under TTPS, TTVS, or TTTS with parameter  $\beta$ , and this is shown to be optimal among a restricted class of allocation rules. In particular, we observe that  $\beta$  controls the fraction of measurement effort allocated to the true best design  $I^*$ , in the sense that  $\bar{\psi}_{n,I^*} \rightarrow \beta$  as  $n \rightarrow \infty$  under each of the proposed algorithms. A lower bound establishes that no algorithm that allocates a fraction  $\beta$  of overall effort to measuring  $I^*$  can 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 the measurement effort is allocated among the  $k - 1$  suboptimal designs in an asymptotically optimal manner. The final part of the theorem shows that  $\Gamma_\beta^*$  is close to the largest possible exponent  $\Gamma^*$  whenever  $\beta$  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.

**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  $\beta^*$ ,

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pi_n(\Theta_{I^*}^c) = \Gamma^*.$$

Under any adaptive allocation rule,

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \Pi_n(\Theta_{I^*}^c) \leq \Gamma^*.$$

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

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log \Pi_n(\Theta_{I^*}^c) = \Gamma_\beta^* \quad \text{and} \quad \lim_{n \rightarrow \infty} \bar{\psi}_{n,I^*} = \beta.$$

Under any adaptive allocation rule,

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \Pi_n(\Theta_{I^*}^c) \leq \Gamma_\beta^* \quad \text{on any sample path with} \quad \lim_{n \rightarrow \infty} \bar{\psi}_{n,I^*} = \beta.$$

3.  $\Gamma^* \leq 2\Gamma_{\frac{1}{2}}^*$  and

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

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 5 and 6. Proposition 7 shows the top-two rules attain these optimal exponents. Part 3 is stated as Lemma 2 in Section 6.

## 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 and Bubeck \[2010\]](#)). This bound depends on the gaps between the means of the different observation distributions.

We say that a real valued random variable  $X$  is  $\sigma$ -sub-Gaussian if  $\mathbf{E}[\exp\{\lambda(X - \mathbf{E}[X])\}] \leq \exp\left\{\frac{\lambda^2\sigma^2}{2}\right\}$  so that the moment generating function of  $X - \mathbf{E}[X]$  is dominated by that of a zero mean Gaussian random variable with variance  $\sigma^2$ . 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 each  $\theta \in [\underline{\theta}, \bar{\theta}]$ , if  $Y \sim p(y|\theta)$ , then  $Y$  is sub-Gaussian with parameter  $\sigma$ . Then*

$$\Gamma_{\frac{1}{2}}^* \geq \frac{1}{16\sigma^2 \sum_{i \neq I^*} \Delta_i^{-2}}$$

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

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

is the difference between the mean under  $\theta_{I^*}^*$  and the mean under  $\theta_i^*$ .

This shows that  $\Pi_n(\Theta_{I^*}^c)$  decays at asymptotic rate faster than  $\exp\{-\frac{n \min_i \Delta_i^2}{16k\sigma^2}\}$ , 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  $\Delta_i$  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 which 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  $i \neq I^*$ ,*

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

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

## 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  $\frac{1}{n} \log(\frac{a_n}{b_n}) \rightarrow 0$  as  $n \rightarrow \infty$ . This notation means that  $a_n$  and  $b_n$  are *equal up to first order in the exponent*. With this notation, Theorem 1 implies the top-two sampling rules with parameter  $\beta$  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 and Bubeck, 2010].

## 6.2 Posterior Consistency

The next proposition provides a consistency and anti-consistency 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  $\theta_i^*$ . 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 3.** *With probability 1, for any  $i \in \{1, \dots, k\}$  if  $\Psi_{n,i} \rightarrow \infty$ , then, for all  $\epsilon > 0$*

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

*If  $\mathcal{I} = \{i \in \{1, \dots, k\} | \lim_{n \rightarrow \infty} \Psi_{n,i} < \infty\}$  is nonempty, then*

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

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

This result is the key to establishing that  $\alpha_{n,I^*} \rightarrow 1$  under each of the proposed algorithm. The next subsection gives a more refined result that allows us to 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  $\theta$  over  $\theta'$ . 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  $\{1, \dots, k\}$ ,

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

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  $i$ ,  $\psi_{n,i} \triangleq \mathbf{P}(I_n = i | \mathcal{F}_{n-1})$ , changes over time as data is collected. Recall that  $\bar{\psi}_{n,i} \triangleq n^{-1} \sum_{\ell=1}^n \psi_{\ell,i}$  captures the fraction of overall effort allocated to measuring design  $i$  over the first  $n$  periods. Under an adaptive allocation rule,  $\bar{\boldsymbol{\psi}}_n$  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 allocation according to  $\bar{\boldsymbol{\psi}}_n$ ,  $D_{\bar{\boldsymbol{\psi}}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta})$  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  $\inf_{\boldsymbol{\theta} \in \tilde{\Theta}} D_{\bar{\boldsymbol{\psi}}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta})$ , which captures the element in  $\tilde{\Theta}$  that is hardest to distinguish from  $\boldsymbol{\theta}^*$  based on samples from  $\bar{\boldsymbol{\psi}}_n$ .

**Proposition 4.** *For any open set  $\tilde{\Theta} \subset \Theta$ ,*

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

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  $n - 1$  i.i.d terms, each of which has mean

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

Therefore, by the law of large numbers, as  $n \rightarrow \infty$ ,  $n^{-1} \log(\pi_n(\boldsymbol{\theta})/\pi_n(\boldsymbol{\theta}^*)) \rightarrow -d(\theta_i^* || \theta_i)$ , 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  $(\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \dots)$  from a fixed probability distribution  $\boldsymbol{\psi}$ , then

$$\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} \doteq \exp \{ -n D_{\boldsymbol{\psi}}(\boldsymbol{\theta}^* || \boldsymbol{\theta}) \}. \quad (6)$$

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

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

These are the main ideas behind Proposition 4, 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  $\bar{\boldsymbol{\psi}}_n$  changes over time, the point  $\arg \min_{\boldsymbol{\theta} \in \tilde{\Theta}} D_{\bar{\boldsymbol{\psi}}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta})$  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 since she does not know  $\theta^*$ , she does not know which measurements will provide the most information. To characterize the optimal exponent  $\Gamma^*$ , however, it is useful to consider the easier problem of gathering the most effective evidence when  $\theta^*$  is known. We can cast this as a game between two players:

- An experimenter, who knows the true parameter  $\theta^*$ , chooses a (possibly adaptive) measurement rule.
- A referee observes the resulting sequence of observations  $(I_1, Y_{1,I_1}, \dots, I_{n-1}, Y_{n-1,I_{n-1}})$  and computes posterior beliefs  $(\alpha_{n,1}, \dots, \alpha_{n,k})$  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  $\alpha_{n,I^*} \rightarrow 1$  as  $n \rightarrow \infty$ .

In order to drive the posterior probability  $\alpha_{n,I^*}$  to 1, the decision-maker must be able to rule out all parameters in  $\Theta_{I^*}^c$  under which the optimal action is not  $I^*$ . Our analysis shows that the posterior probability assigned to  $\Theta_{I^*}^c$  is dominated by the parameter that is hardest to distinguish from  $\theta^*$  under  $\bar{\psi}_n$ . In particular, by Proposition 4,

$$\Pi_n(\Theta_{I^*}^c) \doteq \exp \left\{ -n \left( \min_{\theta \in \Theta_{I^*}^c} D_{\bar{\psi}_n}(\theta^* || \theta) \right) \right\}$$

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

$$\max_{\psi} \min_{\theta \in \Theta_{I^*}^c} D_{\psi}(\theta^* || \theta) \tag{7}$$

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 which 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, since the set of parameter vectors under which  $i$  is optimal

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

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

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

under which the value at  $i$  exceeds that at  $I^*$ . Since, ignoring the boundary of the set,  $\Theta_{I^*}^c = \cup_{i \neq I^*} \bar{\Theta}_i$ ,

$$\max_{i \neq I^*} \Pi_n(\bar{\Theta}_i) \leq \Pi_n(\Theta_{I^*}^c) \leq k \max_{i \neq I^*} \Pi_n(\bar{\Theta}_i)$$

and therefore

$$\Pi_n(\Theta_{I^*}^c) \doteq \max_{i \neq I^*} \Pi_n(\bar{\Theta}_i). \tag{8}$$

This yields an analogue of (7) that will simplify our subsequent analysis. Combining (8) with Proposition 4 shows the solution to the max-min problem

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

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

$$C_i(\beta, \psi) \triangleq \min_{x \in \mathbb{R}} \beta d(\theta_{I^*}^* || x) + \psi d(\theta_i^* || x). \quad (10)$$

The next lemma shows that the function  $C_i$  arises as the solution to the minimization problem over  $\theta \in \bar{\Theta}_i$  in (9). It also shows that the minimum in (10) is attained by a parameter  $\bar{\theta}$  under which the mean observation is a weighted combination of the means under  $\theta_{I^*}^*$  and  $\theta_i^*$ . 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  $\theta$ .

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

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

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

$$C_i(\psi_{I^*}, \psi_i) = \psi_{I^*} d(\theta_{I^*}^* || \bar{\theta}) + \psi_i d(\theta_i^* || \bar{\theta}),$$

*where  $\bar{\theta} \in [\theta_i^*, \theta_{I^*}^*]$  is the unique solution to*

$$A'(\bar{\theta}) = \frac{\psi_{I^*} A'(\theta_{I^*}^*) + \psi_i A'(\theta_i^*)}{\psi_{I^*} + \psi_i}.$$

Lemma 1 and equation (9) immediately imply

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

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  $\beta$  and  $\psi$ . Naturally, it is an increasing function of the measurement effort  $(\beta, \psi)$  allocated to designs  $I^*$  and  $i$ . For given  $\beta$  and  $\psi$ ,  $C_i(\beta, \psi) \geq C_j(\beta, \psi)$  when  $\theta_i^* \leq \theta_j^*$ , reflecting that  $\theta_i^*$  is easier to distinguish from  $\theta_{I^*}^*$  than  $\theta_j^*$ .

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

$$C_i(\beta, \psi_i) = \left( \frac{\beta \psi_i}{\beta + \psi_i} \right) \frac{(\theta_{I^*}^* - \theta_i^*)^2}{2}.$$

*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  $\sigma^2 = 1/\beta + 1/\psi_i = (\beta + \psi_i)/(\beta\psi_i)$ . Standard Gaussian tail bounds imply that as  $n \rightarrow \infty$ ,  $\mathbf{P}(X_{I^*} - X_i < 0) \doteq \exp(-n/2(\sigma\Delta)^2)$ , 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  $\psi_{n,i} \triangleq \mathbf{P}(I_n = i | \mathcal{F}_{n-1})$  denotes the measurement effort assigned design  $i$  at time  $n$ .

**Proposition 5.** *Let  $\psi^*$  denote the optimal solution to the maximization problem (11). If  $\psi_n = \psi^*$  for all  $n$ , then*

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

Moreover under any other adaptive allocation rule,

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \Pi_n(\Theta_{I^*}^c) \leq \Gamma^*.$$

This shows that under the fixed allocation rule  $\psi^*$  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  $\beta$ -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

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

This optimization problem yields the optimal allocation subject to a constraint that  $\beta$ -fraction of the samples are spent on  $I^*$ . The next subsection will show that TTTS, TTPS, and TTVS attain the error exponent  $\Gamma_\beta^*$ . 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 it is optimal to decrease  $\psi_2$  and increase  $\psi_1$ , 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 closely related to those in [Glynn and Juneja \[2004\]](#), in which large deviations rate functions take the place of the functions  $C_i$ .

**Proposition 6.** *The solution to the optimization problem (12) is the unique allocation  $\psi^*$  satisfying  $\psi_{I^*}^* = \beta$  and*

$$C_i(\beta, \psi_i) = C_j(\beta, \psi_j) \quad \forall i, j \neq I^*.$$

If  $\psi_n = \psi^*$  for all  $n$ , then

$$\Pi_n(\Theta_{I^*}^c) \doteq \exp\{-n\Gamma_\beta^*\}.$$

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

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \Pi_n(\Theta_{I^*}^c) \leq \Gamma_\beta^*$$

almost surely.

The following lemma relates the constrained exponent  $\Gamma_\beta^*$  to  $\Gamma^*$ .

**Lemma 2.** 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 Optimal Adaptive Allocation

**Two Stage Procedures.** While the last subsection describes an asymptotically optimal exploration strategy, implementing this strategy requires knowledge of the parameter vector  $\theta^*$ . One simple approach to attaining the rate (11) is to split the experiment into two phases. For the first  $o(n)$  periods the algorithm selects actions uniformly at random, after which it constructs a point estimate  $\hat{\theta}$  of  $\theta^*$ . In the second phase, it solves the optimization problem (11) with  $\hat{\theta}$  in place of  $\theta^*$ , and follows that allocation for the remaining periods. In the design of sequential experiments, this idea dates back to [Kiefer and Sacks \[1963\]](#), who builds on the work of [Chernoff \[1959\]](#). For the problem of best arm identification, it dates back at least to [Jennison et al. \[1982\]](#).

These two-stage rules can be shown to attain the optimal large deviations rates described in the previous section. But they also have substantial practical limitations, which were discussed explicitly in early papers. [Jennison et al. \[1982\]](#) writes 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  $\theta$  based on a few observations with as much respect as that based on many observations.”

**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. While 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  $\Gamma_{\beta}^*$ .

**Proposition 7.** Under the TTTS, TTPS, or TTVS algorithm with parameter  $\beta > 0$ ,  $\bar{\psi}_n \rightarrow \psi^{\beta}$ , where  $\psi^{\beta}$  is the unique allocation with  $\psi_{I^*}^{\beta} = \beta$  satisfying

$$C_i(\beta, \psi_i^{\beta}) = C_j(\beta, \psi_j^{\beta}) \quad \forall i, j \neq I^*.$$

Therefore,

$$\Pi_n(\Theta_{I^*}^c) \doteq e^{-n\Gamma_{\beta}^*}.$$

To understand this result, imagine that  $n$  is very large, and  $\bar{\psi}_{n, I^*} \approx \beta$ . If the algorithm has allocated too much measurement effort to a suboptimal action  $i$ , with  $\bar{\psi}_{n, i} > \psi_i^{\beta} + \delta$  for a constant  $\delta > 0$ , then it must have allocated too little measurement effort to at least one other suboptimal

design  $j \neq i$ . Since much less evidence has been gathered about  $j$  than  $i$ , we expect  $\alpha_{j,n} \gg \alpha_{i,n}$ . When this occurs, TTTS, TTPS and TTVS essentially never sample action  $i$  until the average effort  $\bar{\psi}_{n,i}$  allocated to design  $i$  dips back down toward  $\psi_i^\beta$ . 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 3.** *For any  $i \neq I^*$ ,  $V_{n,i} \doteq \alpha_{n,i}$*

Note that by this lemma,

$$\Pi_n(\Theta_{I^*}^c) = \sum_{i \neq I^*} \alpha_{n,i} \doteq \sum_{i \neq I^*} V_{n,i},$$

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  $V_{n,i} = \int_{\Theta_i} v_i(\boldsymbol{\theta}) \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta}$  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  $n \rightarrow \infty$ , since it tends to zero at an exponential rate in  $n$  whereas  $v_i(\boldsymbol{\theta})$  is a fixed function of  $n$ .

## 7 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.

**Top-Two Sampling Via Constrained MAP Estimation.** Here we present a version of top-two sampling that uses MAP estimation. This can simplify computations, as MAP estimates can be computed without solving for the normalizing 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  $\hat{\boldsymbol{\theta}}' \in \arg \max_{\boldsymbol{\theta} \in \Theta_{\hat{I}_n}^c} \pi_n(\boldsymbol{\theta})$  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, while 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.

**Indifference Zone Criterion.** Suppose our goal is to confidently identify an  $\epsilon$ -optimal arm, for a user specified indifference parameter  $\epsilon > 0$ . Much of the paper investigates the set of parameters  $\Theta_i$  under which arm  $i$  is optimal, and studies the rate at which  $\Pi_n(\Theta_{I^*}) \rightarrow 1$ . Now, let us instead consider the set of parameters

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

under which  $i$  is  $\epsilon$ -optimal. It is easy to develop a variety of modified top-two sampling rules under which  $\max_i \Pi_n(\Theta_{\epsilon,i}) \rightarrow 1$  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  $\epsilon$ -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  $J_n = \arg \max_i \theta_i$ .

**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 Subsection 6.4.

**Extremely Correlated Designs.** While 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  $x_i^T \boldsymbol{\theta}$ . The vector  $\boldsymbol{\theta} \in \mathbb{R}^d$  is unknown, but we begin with a prior  $\boldsymbol{\theta} \sim N(0, I)$  and see noisy observations of  $x_i^T \boldsymbol{\theta}$  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.

**Tuning  $\beta$ .** The most glaring gap in this work may be arbitrary choice of tuning parameter  $\beta$ . 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  $\beta$  automatically through value of information calculations, or avoid the need for such a parameter altogether.

**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  $\tau_\delta$  be a random variable indicating the stopping time under constraint  $\delta$ . Since  $1 - \max_i \alpha_{n,i} \doteq e^{-n\Gamma_\beta^*}$  under top-two sampling, our results imply that for each sample path  $\tau_\delta \sim \Gamma_\beta^* \log(1/\delta)$  as  $\delta \rightarrow 0$ . It is natural to conjecture that  $\mathbf{E}[\tau_\delta] \sim \Gamma_\beta^* \log(1/\delta)$  as well. This closely mirrors optimal results in Chernoff [1959], Jennison et al. [1982] and Kaufmann [2016]. Does this rule also yield a frequentist probability of incorrect selection that is  $O(\delta)$  as  $\delta \rightarrow 0$ ? 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  $\mathbf{E}[\tau_\delta]$  as in Jennison et al. [1982] or Kaufmann [2016].

## References

- S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT)*, 2012.
- A. E. Albert. The sequential design of experiments for infinitely many states of nature. *The Annals of Mathematical Statistics*, pages 774–799, 1961.
- J.-Y. Audibert and S. Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.
- R. E. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, pages 16–39, 1954.
- D. A. Berry. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science*, pages 175–187, 2004.
- S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pages 23–37. Springer, 2009.
- O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems (NIPS)*, 2011.
- C.-H. Chen, J. Lin, E. Yücesan, and S. E. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.
- H. Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, pages 755–770, 1959.
- H. Chernoff. Approaches in sequential design of experiments. *A Survey of Statistical Design and Linear Models (Edited by J. N. Srivastava)*, pages 67–90, 1975.
- S. E. Chick and P. Frazier. Sequential sampling with economics of selection procedures. *Management Science*, 58(3):550–569, 2012.
- S. E. Chick and N. Gans. Economic analysis of simulation selection problems. *Management Science*, 55(3):421–437, 2009.

- S. E. Chick, J. Branke, and C. Schmidt. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.
- E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory*, pages 255–270. Springer, 2002.
- P. I. Frazier. A fully sequential elimination procedure for indifference-zone ranking and selection with tight bounds on probability of correct selection. *Operations Research*, 62(4):926–942, 2014.
- P.I. Frazier, W.B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.
- V. Gabillon, M. Ghavamzadeh, and A. Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.
- Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory (COLT)*, 2016.
- J.C. Gittins. Bandit processes and dynamic allocation indices. *J. R. Statist. Soc. B*, 41:148–164, 1979.
- J.C. Gittins and D.M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, NL, 1974.
- P. Glynn and S. Juneja. A large deviations perspective on ordinal optimization. In *Simulation Conference, 2004. Proceedings of the 2004 Winter*, volume 1. IEEE, 2004.
- P. Glynn and S. Juneja. Ordinal optimization-empirical large deviations rate estimators, and stochastic multi-armed bandits. *arXiv preprint arXiv:1507.04564*, 2015.
- A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *Proceedings of The 31st International Conference on Machine Learning*, pages 100–108, 2014.
- T. Graepel, J.Q. Candela, T. Borchert, and R. Herbrich. Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft’s Bing search engine. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 13–20, 2010.
- R.M. Gray. *Entropy and information theory*. Springer, 2011.
- S. S. Gupta and K. J. Miescke. Bayesian look ahead one-stage sampling allocations for selection of the best population. *Journal of statistical planning and inference*, 54(2):229–244, 1996.
- E. Gutin and V. Farias. Optimistic gittins indices. In *Advances In Neural Information Processing Systems*, pages 2737–2745, 2016.
- K. Jamieson and R. Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–6. IEEE, 2014.
- C. Jennison, I. M. Johnstone, and B. W. Turnbull. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. *Statistical decision theory and related topics III*, 2:55–86, 1982.
- K. Johnson, D. Simchi-Levi, and H. Wang. Online network revenue management using thompson



- sampling. *Available at SSRN*, 2015.
- Z. Karnin, T. Koren, and O. Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1238–1246, 2013.
- E. Kauffmann, N. Korda, and R. Munos. Thompson sampling: an asymptotically optimal finite time analysis. In *International Conference on Algorithmic Learning Theory*, 2012.
- E. Kaufmann. On bayesian index policies for sequential resource allocation. *arXiv preprint arXiv:1601.01190*, 2016.
- E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251, 2013.
- E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian upper confidence bounds for bandit problems. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.
- E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best arm identification in multi-armed bandit models. *arXiv preprint arXiv:1407.4443*, 2014.
- R. Keener. Second order efficiency in the sequential design of experiments. *The Annals of Statistics*, pages 510–532, 1984.
- J. Kiefer and J. Sacks. Asymptotically optimum sequential inference and design. *The Annals of Mathematical Statistics*, pages 705–750, 1963.
- S.-H. Kim and B. L. Nelson. Selecting the best system. *Handbooks in operations research and management science*, 13:501–534, 2006.
- S.-H. Kim and B. L. Nelson. Recent advances in ranking and selection. In *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, pages 162–172. IEEE Press, 2007.
- N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for one-dimensional exponential family bandits. In *Advances in Neural Information Processing Systems*, 2013.
- T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.
- M. Naghshvar, T. Javidi, et al. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013.
- S. Nitinawarat, G. K. Atia, and V. V. Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013.
- E. Paulson. A sequential procedure for selecting the population with the largest mean from  $k$  normal populations. *The Annals of Mathematical Statistics*, pages 174–180, 1964.
- Y. Rinott. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics-Theory and methods*, 7(8):799–811, 1978.

- D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014.
- D. Russo and J. Zou. How much does your data exploration overfit? Controlling bias via information usage. *arXiv preprint arXiv:1511.05219*, 2015.
- I. O. Ryzhov. On the convergence rates of expected improvement methods. *Operations Research*, 2016.
- I.O. Ryzhov, W.B. Powell, and P.I. Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.
- S.L. Scott. Overview of content experiments: Multi-armed bandit experiments, 2016. URL <https://support.google.com/analytics/answer/2844870?hl=en>. [Online; accessed 9-November-2016].
- N. Srinivas, A. Krause, S.M. Kakade, and M. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, may 2012. ISSN 0018-9448. doi: 10.1109/TIT.2011.2182033.
- L. Tang, R. Rosales, A. Singh, and D. Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1587–1594. ACM, 2013.
- W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.
- S. S. Villar, J. Bowden, J. Wason, et al. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199–215, 2015.
- D. Williams. *Probability with martingales*. Cambridge university press, 1991.

## A Outline

This technical appendix is organized as follows.

1. Section B describes a numerical algorithm that can be used to implement TTPS.
2. Section C provides a more precise discussion of related work by Ryzhov [2016].
3. The theoretical analysis begins in Section D. There we begin by noting some basic facts of exponential family distributions, as well as some results relating martingales to their quadratic variation process.
4. Section E establishes results related to the concentration of the posterior distribution, including the proofs of Prop. 3, Prop. 4, and Lemma 3.
5. Section F studies and simplifies the optimal exponents  $\Gamma^*$  and  $\Gamma_\beta^*$ , including the proofs of Lemma 1, Prop. 6, Lemma 2, Prop. 1, and Prop. 2.
6. We conclude with Section G, which studies the top-two allocation rules and provides a proof of Prop. 7.

## B An Implementation of TTPS

This section describes an implementation of the top-two probability sampling for a problem with a Beta prior and binary observations. In this problem, measurements are binary with success probability given by  $\mathbf{P}(Y_{n,i} = 1) = \theta_i^*$ . The algorithm begins with an independent prior, under which the  $i$ th component of  $\boldsymbol{\theta}$  follows a Beta distribution with parameters  $(\lambda_i^1, \lambda_i^2)$ . When  $\lambda_i^1 = \lambda_i^2 = 1$ , this specifies a uniform prior over  $[0, 1]$ . This prior distribution can be easily updated to form a posterior distribution according to the update rule given in line 19 of Algorithm 3.

This algorithm uses quadrature to approximate the integral defining  $\alpha_{n,i}$ . To understand this implementation, consider a random vector  $(X_1, \dots, X_K)$  whose components are independently distributed with  $X_i \sim \text{Beta}(\lambda_i^1, \lambda_i^2)$ . Then, the probability component  $i$  is maximal can be computed according to

$$\begin{aligned} \mathbf{P}(X_i = \max_j X_j) &= \int_{x \in \mathbb{R}} \mathbf{P}(\cap_{j \neq i} \{X_j \leq x\}) \mathbf{P}(X_i = dx) \\ &= \int_{x \in \mathbb{R}} \left[ \prod_{j \neq i} \mathbf{P}(X_j \leq x) \right] \mathbf{P}(X_i = dx) \\ &= \int_{x \in \mathbb{R}} \left[ \left( \prod_{j=1}^K \mathbf{P}(X_j \leq x) \right) / \mathbf{P}(X_i \leq x) \right] \mathbf{P}(X_i = dx). \end{aligned}$$

Algorithm 3 takes as input a vector of  $\mathbf{x}$  consisting of  $M$  points in  $(0, 1)$  and approximates the above integral using quadrature at these points. The algorithm computes and updates the posterior PDF and CDF of  $\theta_i$  in an  $M$  dimensional vectors  $\mathbf{f}_i$  and  $\mathbf{F}_i$ . It also stores and updates a vector  $\bar{\mathbf{F}} = \prod_{i=1}^K \bar{F}_{i,m}$ , where  $\bar{F}_m$  is the posterior probability all the designs have quality below  $x_m$ . Using these quantities, the posterior probability design  $i$  is optimal is approximated by a sum in line 8. Lines 11-15 select an action according to TTPS and lines 18-21 update the stored statistics of the posterior using Bayes rule. The algorithm continues for  $N$  time steps, and upon stopping returns the posterior parameters  $\boldsymbol{\lambda}^1$  and  $\boldsymbol{\lambda}^2$ , which summarize all evidence gathered throughout the measurement process. The algorithm has  $O(NKM)$  space and time complexity. It is worth noting that most operations in this algorithm can be implemented in a “vectorized” fashion in languages like MATLAB, NumPy, and Julia.

---

**Algorithm 3** BernoulliTTPS( $\beta, K, M, N, \lambda^1, \lambda^2, \mathbf{x}$ )

---

```

1: \Initialize:
2:  $f_{i,m} \leftarrow \text{Beta.pdf}(x_m | \lambda_i^1, \lambda_i^2) \quad \forall i, m$ 
3:  $F_{i,m} \leftarrow \text{Beta.cdf}(x_m | \lambda_i^1, \lambda_i^2) \quad \forall i, m$ 
4:  $\bar{F}_m \leftarrow \prod_i F_{i,m} \quad \forall m$ 
5:
6: for  $n = 1 \dots N$  do
7:   \Compute Optimal Action Probabilities:
8:    $\alpha_i \leftarrow \sum_m f_{i,m} \bar{F}_m / F_{i,m} \quad \forall i$ 
9:
10:  \Act and Observe:
11:   $J_1 \leftarrow \arg \max_i \alpha_i$ 
12:   $J_2 \leftarrow \arg \max_{i \neq J_1} \alpha_i$ 
13:  Sample  $B \sim \text{Bernoulli}(\beta)$ 
14:   $I \leftarrow BJ_1 + (1 - B)J_2$ .
15:  Play  $I$  and Observe  $Y_{n,I} \in \{0, 1\}$ .
16:
17:  \Update Statistics:
18:   $(\lambda_I^1, \lambda_I^2) \leftarrow (\lambda_I^1, \lambda_I^2) + (Y_{n,I}, 1 - Y_{n,I})$ 
19:   $\bar{F}_m \leftarrow (\bar{F}_m / F_{I,m}) \times \text{Beta.cdf}(x_m | \lambda_I^1, \lambda_I^2) \quad \forall m$ 
20:   $F_{I,m} \leftarrow \text{Beta.cdf}(x_m | \lambda_I^1, \lambda_I^2) \quad \forall m$ 
21:   $f_{I,m} \leftarrow \text{Beta.pdf}(x_m | \lambda_I^1, \lambda_I^2) \quad \forall m$ 
22: end for
23: return  $V, \lambda^1, \lambda^2$ 

```

---

## C Discussion of the Expected Improvement Algorithm

Here, we briefly discuss interesting recent results of Ryzhov [2016]. He studies a setting with an uncorrelated Gaussian prior, and Gaussian observation noise  $Y_{n,i} \sim N(\theta_i, \sigma_i^2)$ . To simplify our discussion, let us restrict attention to the case of common variance  $\sigma_1 = \dots = \sigma_k = \sigma$ . Ryzhov [2016] shows that under the the expected-improvement algorithm, in the limit as  $n \rightarrow \infty$

$$\sum_{i \neq I^*} \Psi_{n,i} = O(\log n) \quad (13)$$

and

$$\Psi_{n,i}(\theta_I^* - \theta_i)^2 \sim \Psi_{n,j}(\theta_I^* - \theta_j)^2 \quad \forall i, j \neq I^* \quad (14)$$

Recall that  $\Psi_{n,i} = \sum_{\ell=1}^n \psi_{n,i}$  denotes the total measurement effort allocated to design  $i$ . The sampling ratios (14) are the ratios suggested in the optimal computing budget allocation of Chen et al. [2000]. This work therefore establishes an interesting link between EI and OCBA, which appear quite different on the surface.

Unfortunately, property (13) is not suggested by the OCBA, and implies that  $\Pi_n(\Theta_{I^*}^c)$  cannot tend to zero at an exponential rate. To see this precisely, assume without loss of generality that  $I^* = 1$ . Then (13) implies  $\bar{\psi}_n \rightarrow \mathbf{e}_1 \equiv (1, 0, 0, \dots, 0)$ . It is easy to show that  $\min_{\theta \in \Theta_1^c} D_{\mathbf{e}_1}(\theta^* || \theta) = 0$  and therefore, by Proposition 4,

$$\lim_{n \rightarrow \infty} -n^{-1} \log \Pi_n(\Theta_1^c) = 0.$$

It is also worth noting that the sampling ratios in (14) are not actually optimal for any finite number of designs  $k$ . Specifying our calculations as in Example 1, one can show that under an optimal fixed allocation  $(\psi_i, \dots, \psi_k)$ ,

$$\frac{(\theta_{I^*}^* - \theta_i^*)^2}{1/\psi_{I^*} + 1/\psi_i} = \frac{(\theta_{I^*}^* - \theta_j^*)^2}{1/\psi_{I^*} + 1/\psi_j} \quad \forall i, j \neq I^*.$$

These calculations match those in Glynn and Juneja [2004] and Jennison et al. [1982]. As a result, there is no problem with finite  $k$  for which the sampling ratios in (14) are optimal. One can show, in fact, that any optimal multi-armed bandit algorithm that attains the lower bound of Lai and Robbins [1985] also satisfies equations (13) and (14). The main innovation in this paper is to show how to build on such bandit algorithms to attain near-optimal rates for the best-arm identification problem.

Ryzhov [2016] also studies the knowledge gradient policy, which could offer improved performance as (13) no longer holds, but shows that as  $n \rightarrow \infty$

$$\Psi_{n,i}(\theta_{I^*}^* - \theta_i) \sim \Psi_{n,j}(\theta_{I^*}^* - \theta_j) \quad \forall i, j \neq I^*,$$

which could be very far from the optimal sampling proportions.

## D Preliminaries

This section presents some basic results which will be used in the subsequent analysis. First, unless clearly specified, all statements about random variables are meant to hold with probability 1. So for sequences of random variables  $\{X_n\}$  and  $\{Y_n\}$ , if we say that  $X_n \rightarrow \infty$  whenever  $Y_n \rightarrow \infty$ , this means that the set  $\{\omega : Y_n(\omega) \rightarrow \infty, X_n(\omega) \not\rightarrow \infty\}$  has measure zero.

**Facts about the exponential family.** The log partition function  $A(\theta)$  is strictly convex and differentiable, with

$$A'(\theta) = \int T(y)p(y|\theta)d\nu(y) \tag{15}$$

equal to the mean under  $\theta$ . The Kullback-Leibler divergence is equal to

$$d(\theta||\theta') = (\theta - \theta')A'(\theta) - A(\theta) + A(\theta') \tag{16}$$

and satisfies

$$\theta'' > \theta' \geq \theta \implies d(\theta||\theta'') > d(\theta||\theta') \tag{17}$$

$$\theta'' < \theta' \leq \theta \implies d(\theta||\theta'') < d(\theta||\theta'). \tag{18}$$

Finally, since  $[\underline{\theta}, \bar{\theta}]$  is bounded, and we have assumed  $\sup_{\theta \in [\underline{\theta}, \bar{\theta}]} |A'(\theta)| < \infty$ ,

$$\sup_{\theta \in [\underline{\theta}, \bar{\theta}]} |A(\theta)| < \infty \quad \text{and} \quad \sup_{\theta, \theta' \in [\underline{\theta}, \bar{\theta}]} d(\theta||\theta') < \infty. \tag{19}$$

This effectively guarantees no single observation can provide enough information to completely rule out a parameter.

**Some martingale convergence results.** The next fact relates the behavior of a martingale  $M_n$  to its quadratic variation  $\langle M \rangle_n$ .

**Fact 1.** (*Williams [1991], 12.13-12.14*) Let  $\{M_n\}$  be a square-integrable martingale adapted to the filtration  $\{\mathcal{H}_n\}$  and let

$$\langle M \rangle_n = \sum_{\ell=1}^n \mathbf{E}[(M_\ell - M_{\ell-1})^2 | \mathcal{H}_{\ell-1}]$$

denote the corresponding quadratic variation process. Then

$$\frac{M_n}{\langle M \rangle_n} \rightarrow \infty$$

almost surely if  $\langle M \rangle_n \rightarrow \infty$  and  $\lim_{n \rightarrow \infty} M_n$  exists and is finite almost surely if  $\lim_{n \rightarrow \infty} \langle M \rangle_n < \infty$ .

The next lemma is crucial to our analysis. To draw the connection with our setting, imagine an adaptive-randomized rule is used to determine when to draw samples from a population. Here  $Y_n \in \mathbb{R}$  denotes the sample at time  $n$ ,  $X_n \in \{0, 1\}$  indicates whether the sample was measured, and  $Z_n \in [0, 1]$  determines the probability of measurement conditioned on the past. This lemma provides a law of large numbers when measurement effort  $\sum_{\ell=1}^n Z_\ell$  tends to infinity, but shows that if measurement effort is finite then  $\sum_{\ell=1}^{\infty} X_\ell Y_\ell$  is also finite; in this sense the observations collected from  $Y_n$  are inconclusive when measurement effort is finite.

**Lemma 4.** Let  $\{Y_n\}$  be an i.i.d sequence of real-valued random variables with finite variance and let  $\{X_n\}$  be a sequence of binary random variables. Suppose each sequence is adapted to the filtration  $\{\mathcal{H}_n\}$ , and define  $Z_n = \mathbf{P}(X_n = 1 | \mathcal{H}_{n-1})$ . If, conditioned on  $\mathcal{H}_{n-1}$ , each  $Y_n$  is independent of  $X_n$ , then with probability 1,

$$\lim_{n \rightarrow \infty} \sum_{\ell=1}^n Z_\ell = \infty \implies \lim_{n \rightarrow \infty} \frac{\sum_{\ell=1}^n X_\ell Y_\ell}{\sum_{\ell=1}^n Z_\ell} = \mathbf{E}[Y_1]$$

and

$$\lim_{n \rightarrow \infty} \sum_{\ell=1}^n Z_\ell < \infty \implies \sup_{n \in \mathbb{N}} \left| \sum_{\ell=1}^n X_\ell Y_\ell \right| < \infty.$$

*Proof.* Let  $\mu = \mathbf{E}[Y_1]$  and  $\sigma^2 = \mathbf{E}[(Y_1 - \mathbf{E}[Y_1])^2]$  denote the mean and variance of each  $Y_n$ . Define the martingale

$$M_n = \sum_{\ell=1}^n (X_\ell Y_\ell - Z_\ell \mu)$$

with  $M_0 = 0$  and put  $S_n = \sum_{\ell=1}^n Z_\ell$ . This martingale has quadratic variation

$$\begin{aligned} \langle M \rangle_n &= \sum_{\ell=1}^n \mathbf{E}[(M_\ell - M_{\ell-1})^2 | \mathcal{H}_{\ell-1}] \\ &= \sum_{\ell=1}^n \mathbf{E}[(X_\ell(Y_\ell - \mu) + (Y_\ell - Z_\ell)\mu)^2 | \mathcal{H}_{\ell-1}] \\ &= \sum_{\ell=1}^n Z_\ell \sigma^2 + \sum_{\ell=1}^n Z_\ell (1 - Z_\ell) \mu^2 \\ &\leq (\sigma^2 + \mu^2) S_n. \end{aligned}$$

We use the shorthand  $S_\infty = \lim_{n \rightarrow \infty} S_n$  and  $\langle M \rangle_\infty = \lim_{n \rightarrow \infty} \langle M \rangle_n$ .

Suppose  $S_\infty < \infty$  so  $\langle M \rangle_\infty < \infty$ . By Fact 1,  $\lim_{n \rightarrow \infty} M_n$  exists and is finite almost surely, which implies  $\sup_{n \in \mathbb{N}} |M_n| < \infty$ . Since  $|\sum_{\ell=1}^n X_\ell Y_\ell| \leq |M_n| + |\mu S_\infty|$ , this shows  $\sup_{n \in \mathbb{N}} |\sum_{\ell=1}^n X_\ell Y_\ell| < \infty$  as desired.

Now, suppose  $S_\infty = \infty$ . If  $\langle M \rangle_\infty < \infty$ , then again by Fact 1,  $\lim_{n \rightarrow \infty} M_n < \infty$  and it is immediate that  $S_n^{-1} M_n \rightarrow 0$ . However, if  $\langle M \rangle_\infty = \infty$  then

$$\frac{M_n}{\langle M \rangle_n} \rightarrow 0,$$

which implies  $S_n^{-1} M_n \rightarrow 0$  since  $S_n \geq (\sigma^2 + \mu^2) \langle M \rangle_n$ .  $\square$

Taking  $Y_n = 1$  in the lemma above yields Levy's extension of the Borel–Cantelli lemmas (Williams [1991], 12.15). Specialized to our setting, this result relates the long run measurement effort  $\Psi_{n,i} = \sum_{\ell=1}^n \psi_{n,i}$  to the number of times alternative  $i$  is actually measured  $\sum_{\ell=1}^n \mathbf{1}(I_n = i)$ .

**Corollary 1.** *For  $i \in \{1, \dots, k\}$ , set  $S_{n,i} = \sum_{\ell=1}^n \mathbf{1}(I_n = i)$ . Then, with probability 1,*

$$\Psi_{n,i} \rightarrow \infty \iff S_{n,i} \rightarrow \infty$$

and

$$\Psi_{n,i} \rightarrow \infty \implies \frac{S_{n,i}}{\Psi_{n,i}} \rightarrow 1.$$

*Proof.* Apply Lemma 4 with  $Y_n = 1$ ,  $X_n = \mathbf{1}(I_n = 1)$ , and  $\mathcal{H}_n = \mathcal{F}_n$ . Then  $Z_n = \psi_{n,i}$  by definition.  $\square$

## E Posterior Concentration and anti-Concentration

### E.1 Uniform Convergence of the Log-Likelihood

We study the log-likelihood

$$\Lambda_n(\boldsymbol{\theta}^* | \boldsymbol{\theta}) \triangleq \log \left( \frac{L_n(\boldsymbol{\theta}^*)}{L_n(\boldsymbol{\theta})} \right) = \sum_{\ell=1}^n \log \left( \frac{p(Y_{\ell, I_\ell} | \boldsymbol{\theta}_\ell^*)}{p(Y_{\ell, I_\ell} | \boldsymbol{\theta}_\ell)} \right)$$

and the log-likelihood from observations of design  $i$

$$\Lambda_{n,i}(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i) \triangleq \sum_{\ell=1}^n \mathbf{1}(I_n = i) \log \left( \frac{p(Y_{n,i} | \boldsymbol{\theta}_i^*)}{p(Y_{n,i} | \boldsymbol{\theta}_i)} \right).$$

A Doob-decomposition expresses  $\Lambda_{n,i}(\boldsymbol{\theta}_i) = A_n(\boldsymbol{\theta}_i) + M_n(\boldsymbol{\theta}_i)$  as the sum of an  $\mathcal{F}_{n-1}$  predictable process  $A_n(\boldsymbol{\theta}_i)$  and a Martingale  $M_n(\boldsymbol{\theta}_i)$ . Moreover, an easy calculation shows  $A_n(\boldsymbol{\theta}_i) = \Psi_{n,i} d(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i)$  and  $M_n(\boldsymbol{\theta}_i) = \Lambda_{n,i}(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i) - \Psi_{n,i} d(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i)$ . Applying Lemma 4 shows  $\Psi_{n,i}^{-1} M_n(\boldsymbol{\theta}_i) \rightarrow 0$  if  $\Psi_{n,i} \rightarrow \infty$ , which shows the log-likelihood ratio tends to infinity at rate  $\Psi_{n,i} d(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i)$ . The next lemma strengthens this, and provides a link between these quantities that holds uniformly in  $\boldsymbol{\theta}_i$ .

**Lemma 5.** *With probability 1, if  $\Psi_{n,i} \rightarrow \infty$  then*

$$\sup_{\boldsymbol{\theta}_i \in [\underline{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}]} \Psi_{n,i}^{-1} |\Lambda_{n,i}(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i) - \Psi_{n,i} d(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i)| \rightarrow 0,$$

and if  $\lim_{n \rightarrow \infty} \Psi_{n,i} < \infty$  then

$$\sup_{\boldsymbol{\theta}_i \in [\underline{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}]} \sup_{n \in \mathbb{N}} |\Lambda_{n,i}(\boldsymbol{\theta}_i) + |\Psi_{n,i} d(\boldsymbol{\theta}_i^* | \boldsymbol{\theta}_i)| < \infty.$$

*Proof.* Define  $\xi_n \triangleq T(Y_{n,i}) - \mathbf{E}[T(Y_{n,i})]$  and  $X_n \triangleq \mathbf{1}(I_n = i)$ . Note that  $\mathbf{E}[\xi_n | \mathcal{F}_{n-1}] = 0$ ,  $\mathbf{E}[X_n | \mathcal{F}_{n-1}] = \psi_{n,i}$ , and, conditioned on  $\mathcal{F}_{n-1}$ ,  $X_n$  is independent of  $\xi_n$ . Using the form of the exponential family density given in equation (1), and the form of the KL-divergence given in equation (16), the log-likelihood ratio can be written as

$$\begin{aligned} \log \left( \frac{p(Y_{n,i} | \theta_i^*)}{p(Y_{n,i} | \theta_i)} \right) &= (\theta_i^* - \theta_i) T(Y_{n,i}) - (A(\theta_i^*) - A(\theta_i)) \\ &= d(\theta_i^* || \theta_i) + (\theta_i^* - \theta_i) (T(Y_{n,i}) - \mathbf{E}[T(Y_{n,i})]) \\ &= d(\theta_i^* || \theta_i) + (\theta_i^* - \theta_i) \xi_n \end{aligned}$$

Therefore,

$$\begin{aligned} \Lambda_{n,i}(\theta_i^* || \theta_i) - \Psi_{n,i} d(\theta_i^* || \theta_i) &= \sum_{\ell=1}^n X_\ell \log \left( \frac{p(Y_{\ell,i} | \theta_i^*)}{p(Y_{\ell,i} | \theta_i)} \right) - \sum_{\ell=1}^n \psi_{\ell,i} d(\theta_i^* || \theta_i) \\ &= \sum_{\ell=1}^n (X_\ell - \psi_{\ell,i}) d(\theta_i^* || \theta_i) + \sum_{\ell=1}^n X_\ell \xi_\ell (\theta_i^* - \theta_i). \end{aligned}$$

Here  $|\theta_i^* - \theta_i| \leq \bar{\theta} - \underline{\theta} \equiv C_2$  is bounded uniformly. Similarly, as shown in Appendix D,  $d(\theta_i^* || \theta_i)$  is bounded uniformly in  $\theta_i$  by

$$C_1 \equiv \max_{\theta' \in [\underline{\theta}, \bar{\theta}]} d(\theta_i^* || \theta') < \infty.$$

This implies,

$$|\Lambda_{n,i}(\theta_i) - \Psi_{n,i} d(\theta_i^* || \theta_i)| \leq C_1 \left| \sum_{\ell=1}^n (X_\ell - \psi_{\ell,i}) \right| + C_2 \left| \sum_{\ell=1}^n X_\ell \xi_\ell \right| \quad (20)$$

$$|\Lambda_{n,i}(\theta_i)| \leq C_1 \Psi_{n,i} + C_1 \left| \sum_{\ell=1}^n (X_\ell - \psi_{\ell,i}) \right| + C_2 \left| \sum_{\ell=1}^n X_\ell \xi_\ell \right|. \quad (21)$$

Since  $\mathbf{E}[\xi_n^2] < \infty$ , the result then follows by applying Lemma 4 and Corollary 1. In particular, when  $\Psi_{n,i} \rightarrow \infty$ ,

$$\lim_{n \rightarrow \infty} \Psi_{n,i}^{-1} \sum_{\ell=1}^n (X_\ell - \psi_{\ell,i}) = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \Psi_{n,i}^{-1} \sum_{\ell=1}^n X_\ell \xi_\ell = 0$$

When  $\lim_{n \rightarrow \infty} \Psi_{n,i} < \infty$ ,

$$\sup_{n \in \mathbb{N}} \left| \sum_{\ell=1}^n (X_\ell - \psi_{\ell,i}) \right| < \infty \quad \text{and} \quad \sup_{n \in \mathbb{N}} \left| \sum_{\ell=1}^n X_\ell \xi_\ell \right| < \infty.$$

It is also immediate that  $d(\theta_i^* || \theta_i) \Psi_{n,i} \leq C_1 \Psi_{n,i} \not\rightarrow \infty$ , which by (21) implies the second part of the result.  $\square$

A corollary of the previous lemma relates the log-likelihood ratio  $\Lambda_n(\boldsymbol{\theta}^* || \boldsymbol{\theta})$  to the Kullback-Leibler divergence  $D_{\bar{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta})$ .

**Corollary 2.** *With probability 1,*

$$\sup_{\boldsymbol{\theta} \in \Theta} |n^{-1} \Lambda_n(\boldsymbol{\theta}^* || \boldsymbol{\theta}) - D_{\bar{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta})| \rightarrow 0$$



*Proof.*

$$\begin{aligned} \left| n^{-1} \Lambda_n(\boldsymbol{\theta}^* | \boldsymbol{\theta}) - D_{\bar{\psi}_n}(\boldsymbol{\theta}^* | \boldsymbol{\theta}) \right| &= \left| n^{-1} \sum_{i=1}^k (\Lambda_{n,i}(\theta_i^* | \theta_i) - \Psi_{n,i} d(\theta_i^* | \theta_i)) \right| \\ &\leq \sum_{i=1}^k n^{-1} |\Lambda_{n,i}(\theta_i^* | \theta_i) - \Psi_{n,i} d(\theta_i^* | \theta_i)|. \end{aligned}$$

Lemma 5 implies

$$\sup_{\theta_i \in [\underline{\theta}, \bar{\theta}]} n^{-1} |\Lambda_{n,i}(\theta_i^* | \theta_i) - \Psi_{n,i} d(\theta_i^* | \theta_i)| \rightarrow 0,$$

which completes the proof.  $\square$

## E.2 Posterior Consistency: Proof of Prop. 3

**Proposition 3.** *For any  $i \in \{1, \dots, k\}$  if  $\Psi_{n,i} \rightarrow \infty$ , then, for all  $\epsilon > 0$*

$$\Pi_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \notin (\theta_i^* - \epsilon, \theta_i^* + \epsilon)\}) \rightarrow 0,$$

*with probability 1. If  $\mathcal{I} = \{i \in \{1, \dots, k\} | \lim_{n \rightarrow \infty} \Psi_{n,i} < \infty\}$  is nonempty, then*

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

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

Because we don't assume an independent prior across the designs,  $\Pi_1$  is not a product measure and therefore neither is  $\Pi_n$ . This makes it challenging to reason about the marginal posterior of each design, which is required for Proposition 3. Thankfully, since the prior density is bounded,  $\Pi_n$  behaves like a product measure. Note that the likelihood function can be written as the product of  $k$  terms:

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^k L_{n,i}(\theta_i)$$

where

$$L_{n,i}(\theta_i) \triangleq \prod_{\substack{\ell \leq n \\ I_\ell = i}} p(Y_{\ell,1} | \theta_i)$$

with the convention that  $L_{n,i}(\theta_i) = 1$  when  $\sum_{\ell=1}^n \mathbf{1}(I_\ell = i) = 0$ . Therefore  $L_n(\boldsymbol{\theta})$  forms the density of a product measure. By normalizing, this induces a probability measure over  $\Theta$ ,

$$\mathcal{L}_n(\tilde{\Theta}) \triangleq \frac{\int_{\tilde{\Theta}} L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}} \quad \tilde{\Theta} \subset \Theta,$$

which, as we argue in the next lemma, behaves like the posterior  $\Pi_n$ .

**Lemma 6.** *For any set  $\tilde{\Theta} \subset \Theta$ ,*

$$C^{-1} \mathcal{L}_n(\tilde{\Theta}) \leq \Pi_{n+1}(\tilde{\Theta}) \leq C \mathcal{L}_n(\tilde{\Theta}),$$

*where*

$$C = \frac{\sup_{\boldsymbol{\theta} \in \Theta} \pi_1(\boldsymbol{\theta})}{\inf_{\boldsymbol{\theta} \in \Theta} \pi_1(\boldsymbol{\theta})} < \infty$$

*is independent of  $n$  and  $\tilde{\Theta}$ .*

*Proof.* This follows immediately by bounding  $\pi_1(\boldsymbol{\theta})$  from above and below in the relation

$$\Pi_{n+1}(\tilde{\Theta}) = \frac{\int_{\tilde{\Theta}} \pi_1(\boldsymbol{\theta}) L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} \pi_1(\boldsymbol{\theta}) L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

□

We can now prove Proposition 3.

*Proof of Proposition 3.* We begin with the first part of the result. For simplicity of notation, we focus on the upper interval  $\tilde{\Theta} = \{\boldsymbol{\theta} \in \Theta : \theta_i > \theta_i^* + \epsilon\}$ , but results follow identically for the lower interval. We want to show  $\Pi_n(\tilde{\Theta}) \rightarrow 0$ , which occurs if and only if  $\mathcal{L}_n(\tilde{\Theta}) \rightarrow 0$ . Since  $\mathcal{L}_n$  is a product measure,

$$\mathcal{L}_n(\tilde{\Theta}) = \frac{\int_{\theta_i^* + \epsilon}^{\bar{\theta}} L_{n,i}(\theta) d\theta}{\int_{\underline{\theta}}^{\bar{\theta}} L_{n,i}(\theta) d\theta} = \frac{\int_{\theta_i^* + \epsilon}^{\bar{\theta}} (L_{n,i}(\theta)/L_{n,i}(\theta_i^*)) d\theta}{\int_{\underline{\theta}}^{\bar{\theta}} (L_{n,i}(\theta)/L_{n,i}(\theta_i^*)) d\theta} = \frac{\int_{\theta_i^* + \epsilon}^{\bar{\theta}} \exp\{-\Lambda_{n,i}(\theta_i^*|\theta)\} d\theta}{\int_{\underline{\theta}}^{\bar{\theta}} \exp\{-\Lambda_{n,i}(\theta_i^*|\theta)\} d\theta} \quad (22)$$

where  $\Lambda_{n,i}(\theta_i^*|\theta) = \log(L_{n,i}(\theta_i^*)/L_{n,i}(\theta))$ . By Lemma 5, with probability 1 there is a sequence  $a_n \rightarrow 0$  such that  $|\Lambda_{n,i}(\theta_i^*|\theta) - \Psi_{n,i}d(\theta_i^*|\theta)| \leq a_n$  for all  $\theta$ . Then, for  $b_n = e^{a_n}/e^{-a_n} \rightarrow 1$ , one has

$$\mathcal{L}_n(\tilde{\Theta}) \leq \frac{b_n \int_{\theta_i^* + \epsilon}^{\bar{\theta}} \exp\{-\Psi_{n,i}d(\theta_i^*|\theta)\} d\theta}{\int_{\underline{\theta}}^{\bar{\theta}} \exp\{-\Psi_{n,i}d(\theta_i^*|\theta)\} d\theta} \leq \frac{b_n \int_{\theta_i^* + \epsilon}^{\bar{\theta}} \exp\{-\Psi_{n,i}d(\theta_i^*|\theta)\} d\theta}{\int_{\theta_i^* + \epsilon/2}^{\bar{\theta}} \exp\{-\Psi_{n,i}d(\theta_i^*|\theta)\} d\theta}.$$

The integral in the numerator is upper bounded by  $(\bar{\theta} - \theta_i^* - \epsilon) \exp\{-\Psi_{n,i}d(\theta_i^*|\theta_i^* + \epsilon)\}$  while the integral in the denominator is lower bounded by  $(\epsilon/2) \exp\{-\Psi_{n,i}d(\theta_i^*|\theta_i^* + \epsilon/2)\}$ . This shows

$$\mathcal{L}_n(\tilde{\Theta}) \leq c_0 b_n \exp\{-\Psi_{n,i}(d(\theta_i^*|\theta_i^* + \epsilon) - d(\theta_i^*|\theta_i^* + \epsilon/2))\} \rightarrow 0$$

where  $c_0 = 2\epsilon^{-1}(\bar{\theta} - \theta_i^* - \epsilon)$ .

The second part of the claim follows from the lower bound in Lemma 6 of

$$\Pi_{n+1}(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta'_i, \theta''_i) \forall i \in \mathcal{I}\}) \geq C^{-1} \mathcal{L}_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta'_i, \theta''_i) \forall i \in \mathcal{I}\}) \quad (23)$$

$$= C^{-1} \prod_{i \in \mathcal{I}} \mathcal{L}_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta'_i, \theta''_i)\}). \quad (24)$$

As in (22),

$$\mathcal{L}_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta'_i, \theta''_i)\}) = \frac{\int_{\theta'_i}^{\theta''_i} \exp\{-\Lambda_{n,i}(\theta_i^*|\theta)\} d\theta}{\int_{\underline{\theta}}^{\bar{\theta}} \exp\{-\Lambda_{n,i}(\theta_i^*|\theta)\} d\theta}.$$

When  $\lim_{n \rightarrow \infty} \Psi_{n,i} < \infty$ , Lemma 5 shows that for each  $i \in \mathcal{I}$ ,

$$\sup_{\theta_i \in [\underline{\theta}, \bar{\theta}]} \sup_{n \in \mathbb{N}} |\Lambda_{n,i}(\theta_i^*|\theta_i)| < \infty.$$

This implies

$$\inf_n \mathcal{L}_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta'_i, \theta''_i)\}) > 0$$

and establishes the claim. □

### E.3 Large Deviations: Proof of Proposition 4

All statements in this section hold when observations are drawn under the parameter  $\boldsymbol{\theta}^*$ . Since  $\boldsymbol{\theta}^*$  is fixed throughout, we simplify notation and write

$$W_n(\boldsymbol{\theta}) \triangleq D_{\bar{\psi}_n}(\boldsymbol{\theta}^*||\boldsymbol{\theta}).$$

Note that  $W_n(\boldsymbol{\theta}^*) = 0$ . As shown in the next lemma  $n^{-1} \log(\pi_n(\boldsymbol{\theta})/\pi_n(\boldsymbol{\theta}^*)) - W_n(\boldsymbol{\theta}) \rightarrow 0$  uniformly in  $\boldsymbol{\theta}$ .

**Lemma 7.** *With probability 1,*

$$\sup_{\boldsymbol{\theta} \in \Theta} n^{-1} \left| \log \left( \frac{\pi_n(\boldsymbol{\theta}^*)}{\pi_n(\boldsymbol{\theta})} \right) - W_n(\boldsymbol{\theta}) \right| \rightarrow 0.$$

*Proof.* We have

$$\log \left( \frac{\pi_n(\boldsymbol{\theta}^*)}{\pi_n(\boldsymbol{\theta})} \right) - W_n(\boldsymbol{\theta}) = \log \left( \frac{\pi_1(\boldsymbol{\theta}^*)}{\pi_1(\boldsymbol{\theta})} \right) + (\Lambda_{n-1}(\boldsymbol{\theta}^*||\boldsymbol{\theta}) - W_{n-1}(\boldsymbol{\theta})) + (W_{n-1}(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta})).$$

Since  $\inf_{\boldsymbol{\theta} \in \Theta} \pi_1(\boldsymbol{\theta}) > 0$  and  $\sup_{\boldsymbol{\theta} \in \Theta} \pi_1(\boldsymbol{\theta}) < \infty$ ,  $n^{-1} \log(\pi_1(\boldsymbol{\theta})/\pi_1(\boldsymbol{\theta}^*)) \rightarrow 0$  uniformly in  $\boldsymbol{\theta}$ . By Corollary 2,  $n^{-1}(\Lambda_{n-1}(\boldsymbol{\theta}) - W_{n-1}(\boldsymbol{\theta})) \rightarrow 0$  uniformly as well. Finally, by equation (19),  $n^{-1}(W_n(\boldsymbol{\theta}) - W_{n-1}(\boldsymbol{\theta})) \leq n^{-1} \max_i d(\boldsymbol{\theta}_i^*||\boldsymbol{\theta}_i) \rightarrow 0$  uniformly in  $\boldsymbol{\theta}$ .  $\square$

The remaining proof of Proposition 4 follows from a sequence of lemmas. The next observes a form of uniform continuity of  $W_n$  that follows from the uniform bound on  $A'(\boldsymbol{\theta})$  in Assumption 1.

**Lemma 8.** *For all  $\epsilon > 0$ , there exists  $\delta > 0$  such that for  $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty \leq \delta \implies \sup_{n \in \mathbb{N}} |W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}')| \leq \epsilon.$$

*Proof.* We have that

$$\begin{aligned} |W_n(\boldsymbol{\theta}) - W_n(\boldsymbol{\theta}')| &\leq \max_{1 \leq i \leq k} |d(\boldsymbol{\theta}_i^*||\boldsymbol{\theta}_i) - d(\boldsymbol{\theta}_i^*||\boldsymbol{\theta}'_i)| \\ &= \max_{1 \leq i \leq k} |(\boldsymbol{\theta}'_i - \boldsymbol{\theta}_i)A'(\boldsymbol{\theta}_i^*) + A(\boldsymbol{\theta}_i) - A(\boldsymbol{\theta}'_i)| \\ &\leq 2C\delta \end{aligned}$$

where  $C = \sup_{\boldsymbol{\theta} \in (\underline{\theta}, \bar{\theta})} |A'(\boldsymbol{\theta})| < \infty$ .  $\square$

**Lemma 9.** *For any open set  $\tilde{\Theta} \subset \Theta$ ,*

$$\int_{\boldsymbol{\theta} \in \tilde{\Theta}} \frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} d\boldsymbol{\theta} \doteq \int_{\boldsymbol{\theta} \in \tilde{\Theta}} \exp\{-nW_n(\boldsymbol{\theta})\} d\boldsymbol{\theta}.$$

*Proof.* By Corollary 2, we can fix a sequence  $\epsilon_n \geq 0$  with  $\epsilon_n \rightarrow 0$  such that,

$$\exp\{-n(W_n(\boldsymbol{\theta}) + \epsilon_n)\} \leq \frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} \leq \exp\{-n(W_n(\boldsymbol{\theta}) - \epsilon_n)\}.$$

Integrating over  $\tilde{\Theta}$  yields,

$$\exp\{-n\epsilon_n\} \int_{\tilde{\Theta}} \exp\{-nW_n(\boldsymbol{\theta})\} d\boldsymbol{\theta} \leq \int_{\tilde{\Theta}} \frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} d\boldsymbol{\theta} \leq \exp\{n\epsilon_n\} \int_{\tilde{\Theta}} \exp\{-nW_n(\boldsymbol{\theta})\} d\boldsymbol{\theta}.$$

Taking the logarithm of each side implies

$$\frac{1}{n} \left| \log \int_{\tilde{\Theta}} \frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} d\boldsymbol{\theta} - \log \int_{\tilde{\Theta}} \exp\{-nW_n(\boldsymbol{\theta})\} d\boldsymbol{\theta} \right| \leq \epsilon_n \rightarrow 0.$$

□

**Lemma 10.** For any open set  $\tilde{\Theta} \subset \Theta$ ,

$$\int_{\boldsymbol{\theta} \in \tilde{\Theta}} \exp\{-nW_n(\boldsymbol{\theta})\} d\boldsymbol{\theta} \doteq \exp\{-n \inf_{\boldsymbol{\theta} \in \tilde{\Theta}} W_n(\boldsymbol{\theta})\}$$

*Proof.* Let  $\hat{\boldsymbol{\theta}}_n$  be a point in the closure of  $\tilde{\Theta}$ , satisfying

$$W_n(\hat{\boldsymbol{\theta}}_n) = \inf_{\boldsymbol{\theta} \in \tilde{\Theta}} W_n(\boldsymbol{\theta}).$$

Such a point always exists, since  $W_n$  is continuous, and the closure of  $\tilde{\Theta}$  is compact. Let

$$\gamma_n \triangleq \int_{\boldsymbol{\theta} \in \tilde{\Theta}} \exp\{-nW_n(\boldsymbol{\theta})\} d\boldsymbol{\theta}.$$

Our goal is to show

$$\frac{1}{n} \log(\gamma_n) + W_n(\hat{\boldsymbol{\theta}}_n) \rightarrow 0.$$

We have

$$\gamma_n \leq \text{Vol}(\tilde{\Theta}) \exp\{-nW_n(\hat{\boldsymbol{\theta}}_n)\}$$

where for any  $\Theta' \subset \Theta$ ,  $\text{Vol}(\Theta') = \int_{\tilde{\Theta}} d\boldsymbol{\theta} \in (0, \infty)$  denotes the volume of  $\Theta$ . This shows

$$\limsup_{n \rightarrow \infty} \left( \frac{1}{n} \log(\gamma_n) + W_n(\hat{\boldsymbol{\theta}}_n) \right) \leq 0.$$

We now show the reverse. Fix an arbitrary  $\epsilon > 0$ . By Lemma 8, there exists  $\delta > 0$  such that

$$|W_n(\boldsymbol{\theta}) - W_n(\hat{\boldsymbol{\theta}}_n)| \leq \epsilon \quad \forall n \in \mathbb{N}$$

for any  $\boldsymbol{\theta} \in \Theta$  with

$$\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n\|_{\infty} \leq \delta.$$

Now, choose a finite  $\delta$ -cover  $O$  of  $\tilde{\Theta}$  in the norm  $\|\cdot\|_{\infty}$ . Remove any set in  $O$  that does not intersect  $\tilde{\Theta}$ . Then, for each  $o \in O$ ,

$$\text{Vol}(o \cap \tilde{\Theta}) > 0 \implies C_{\delta} \triangleq \min_{o \in O} \text{Vol}(o \cap \tilde{\Theta}) > 0.$$

Choose  $o_n \in O$  with  $\hat{\boldsymbol{\theta}}_n \in \text{closure}(o_n)$ . Then, for every  $\boldsymbol{\theta} \in o_n$ ,  $W_n(\boldsymbol{\theta}) \leq W_n(\hat{\boldsymbol{\theta}}_n) + \epsilon$ . This shows

$$\gamma_n \geq \int_{o_n} \exp\{-nW_n(\boldsymbol{\theta})\} d\boldsymbol{\theta} \geq C_{\delta} \exp\{-n(W_n(\hat{\boldsymbol{\theta}}_n) + \epsilon)\}.$$

Taking the logarithm of both sides implies

$$\frac{1}{n} \log(\gamma_n) + W_n(\hat{\boldsymbol{\theta}}_n) \geq \frac{C_\delta}{n} - \epsilon \rightarrow -\epsilon.$$

Since  $\epsilon$  was chosen arbitrarily, this shows

$$\liminf_{n \rightarrow \infty} \left( \frac{1}{n} \log(\gamma_n) + W_n(\hat{\boldsymbol{\theta}}_n) \right) \geq 0,$$

and completes the proof.  $\square$

We now complete the proof of Proposition 4.

*Proof of Proposition 4.* We begin with a simple observation. For any sequences of real numbers  $\{a_n\}$ ,  $\{b_n\}$ , and  $\{\tilde{a}_n\}$ ,  $\{\tilde{b}_n\}$ , if  $a_n \doteq \tilde{a}_n$  and  $b_n \doteq \tilde{b}_n \in \mathbb{R}$ , then  $a_n/b_n \doteq \tilde{a}_n/\tilde{b}_n$ .

Therefore, we have

$$\Pi_n(\tilde{\Theta}) = \frac{\Pi_n(\tilde{\Theta})}{\Pi_n(\Theta)} = \frac{\int_{\tilde{\Theta}} \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\int_{\tilde{\Theta}} (\pi_n(\boldsymbol{\theta})/\pi_n(\boldsymbol{\theta}^*)) d\boldsymbol{\theta}}{\int_{\Theta} (\pi_n(\boldsymbol{\theta})/\pi_n(\boldsymbol{\theta}^*)) d\boldsymbol{\theta}} \doteq \frac{\exp\{-n \inf_{\boldsymbol{\theta} \in \tilde{\Theta}} W_n(\boldsymbol{\theta})\}}{\exp\{-n \inf_{\boldsymbol{\theta} \in \Theta} W_n(\boldsymbol{\theta})\}}$$

where the final equality follows from the previous two lemmas. Since  $W_n(\boldsymbol{\theta}) \geq 0$  and  $W_n(\boldsymbol{\theta}^*) = 0$ ,  $\exp\{-n \inf_{\boldsymbol{\theta} \in \Theta} W_n(\boldsymbol{\theta})\} = 1$ .  $\square$

## E.4 Large Deviations of the Value Measure: Proof of Lemma 3

**Lemma 3.** For any  $i \neq I^*$ ,  $V_{n,i} \doteq \alpha_{n,i}$ .

*Proof.* First, since

$$V_{n,i} = \int_{\Theta_i} v_i(\boldsymbol{\theta}) \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \leq (u(\bar{\theta}) - u(\underline{\theta})) \int_{\Theta_i} \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta} = (u(\bar{\theta}) - u(\underline{\theta})) \alpha_{n,i}$$

it is immediate that

$$\limsup_{n \rightarrow \infty} n^{-1} (\log V_{n,i} - \log \alpha_{n,i}) \leq 0. \quad (25)$$

The other direction more subtle. Define  $\Theta_{i,\delta} \subset \Theta_i$  by

$$\Theta_{i,\delta} = \{\boldsymbol{\theta} \in \Theta : \theta_i \geq \max_{j \neq i} \theta_j + \delta\}.$$

For any  $\boldsymbol{\theta} \in \Theta_{i,\delta}$ ,  $v_i(\boldsymbol{\theta}) \geq C_\delta$  where

$$C_\delta \equiv \min_{\theta \in [\underline{\theta}, \bar{\theta}]} u(\theta + \delta) - u(\theta) > 0.$$

Because  $u(\theta + \delta) - u(\theta)$  is continuous and is strictly positive for each  $\theta$ , this minimum exists and the objective value is strictly positive. Then

$$V_{n,i} \geq \int_{\Theta_{i,\delta}} v_i(\boldsymbol{\theta}) \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \geq C_\delta \int_{\Theta_{i,\delta}} \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta} = C_\delta \Pi_n(\Theta_{i,\delta}) \quad \forall \delta > 0.$$

Combining this with Proposition 4 shows

$$\liminf_{n \rightarrow \infty} \frac{1}{n} (\log V_{n,i} - \log \alpha_{n,i}) \geq \liminf_{n \rightarrow \infty} \frac{1}{n} (\log \Pi_n(\Theta_{i,\delta}) - \log \Pi_n(\Theta_i)) = - \min_{\boldsymbol{\theta} \in \Theta_{i,\delta}} D_{\bar{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta}) - \min_{\boldsymbol{\theta} \in \Theta_i} D_{\bar{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta}).$$

The final term can be made arbitrarily small by taking  $\delta \rightarrow 0$ . Precisely, by Lemma 8, for any  $\epsilon > 0$ , there exists  $\delta > 0$  such that for all  $n \in \mathbb{N}$  and  $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$  satisfying  $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty \leq \delta$ ,

$$D_{\bar{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta}) \leq \epsilon.$$

Therefore, for each  $\epsilon > 0$  one can choose  $\delta > 0$  such that

$$\min_{\boldsymbol{\theta} \in \Theta_{i,\delta}} D_{\bar{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta}) \leq \min_{\boldsymbol{\theta} \in \Theta_i} D_{\bar{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta}) + \epsilon.$$

This shows  $\liminf n^{-1} (\log V_{n,i} - \log \alpha_{n,i}) \geq -\epsilon$  for all  $\epsilon > 0$ , and hence

$$\liminf_{n \rightarrow \infty} n^{-1} (\log V_{n,i} - \log \alpha_{n,i}) \geq 0.$$

□

## F Simplifying and Bounding the Error Exponent

### F.1 Proof of Lemma 1

To begin, we restate the results of Lemma 1 in the order in which they will be proved. Recall, from Section D that  $A(\theta)$  is increasing and strictly convex, and, by (15),  $A'(\theta)$  is the mean observation under  $\theta$ .

**Lemma 1.** Define for each  $i \neq I^*, \psi \geq 0$ ,

$$C_i(\beta, \psi) \triangleq \min_{x \in \mathbb{R}} \beta d(\theta_{I^*}^* || x) + \psi d(\theta_i^* || x). \quad (26)$$

(a) For any  $i \neq I^*$  and probability distribution  $\psi$  over  $\{1, \dots, k\}$

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

where  $\bar{\Theta}_i \triangleq \{\boldsymbol{\theta} \in \Theta | \theta_i \geq \theta_{I^*}\}$ .

(b) Each  $C_i$  is a concave function.

(c) The unique solution to the minimization problem (26) is  $\bar{\theta} \in \mathbb{R}$  satisfying

$$A'(\bar{\theta}) = \frac{\psi_{I^*} A'(\theta_{I^*}^*) + \psi_i A'(\theta_i^*)}{\psi_{I^*} + \psi_i}.$$

Therefore,

$$C_i(\psi_{I^*}, \psi_i) = \psi_{I^*} d(\theta_{I^*}^* || \bar{\theta}) + \psi_i d(\theta_i^* || \bar{\theta}).$$

(d) Each  $C_i$  is a strictly increasing function.

*Proof.* (a)

$$\begin{aligned}
\min_{\theta \in \Theta_i} D_\psi(\theta^* || \theta) &= \min_{\theta \in \Theta: \theta_i \geq \theta_{I^*}} \sum_{j=1}^k \psi_{n,j} d(\theta_j^* || \theta_j) \\
&= \min_{\bar{\theta} \geq \theta_i \geq \theta_{I^*} \geq \theta} \psi_{I^*} d(\theta_{I^*}^* || \theta_{I^*}) + \psi_i d(\theta_i^* || \theta_i) + \sum_{j \notin \{i, I^*\}} \min_{\theta_j} \psi_{n,j} d(\theta_j^* || \theta_j) \\
&= \min_{\bar{\theta} \geq \theta_i \geq \theta_{I^*} \geq \theta} \psi_{I^*} d(\theta_{I^*}^* || \theta_{I^*}) + \psi_i d(\theta_i^* || \theta_i)
\end{aligned}$$

where the last equality uses that the minimum occurs when  $\theta_j = \theta_j^*$  for  $j \notin \{I^*, i\}$ , and this is feasible for any choice of  $(\theta_i, \theta_{I^*})$ . Then, by the monotonicity properties of KL-divergence (see Section D, equation (17)), there is always a minimum with  $\theta_i = \theta_{I^*}$ . Therefore this objective value is equal to

$$\min_{\theta \in [\underline{\theta}, \bar{\theta}]} \psi_{I^*} d(\theta_{I^*}^* || \theta) + \psi_i d(\theta_i^* || \theta) = \min_{x \in \mathbb{R}} \psi_{I^*} d(\theta_{I^*}^* || x) + \psi_i d(\theta_i^* || x) = C_i(\psi_{I^*}, \psi_i).$$

(b)  $C_i$  is the minimum over a family of linear functions and therefore is concave (See Chapter 3.2 of [Boyd and Vandenberghe \[2004\]](#)). In particular  $C_i(\beta, \psi) = \min_{x \in \mathbb{R}} g((\beta, \psi); x)$  where  $g((\beta, \psi); x) = \beta d(\theta_{I^*}^* || x) + \psi d(\theta_i^* || x)$  is linear in  $(\beta, \psi)$ .

(c) Direct calculation using the formula for KL divergence in exponential families (see (16) in Section D) shows

$$\beta d(\theta_{I^*}^* || x) + \psi_i d(\theta_i^* || x) = (\beta + \psi_i) A(x) - (\beta A'(\theta_{I^*}^*) + \psi_i A'(\theta_i^*)) x + f(\beta, \theta_{I^*}^*, \psi_i, \theta_i^*)$$

where  $f(\beta, \theta_{I^*}^*, \psi_i, \theta_i^*)$  captures terms that are independent of  $x$ . Setting the derivative with respect to  $x$  to zero yields the result since  $A(x)$  is strictly convex.

(d) We will show  $C_i$  is strictly increasing in the second argument. The proof that it is strictly increasing in its first argument follows by symmetry. Set

$$f(\psi_i, x) = \beta d(\theta_{I^*}^* || x) + \psi_i d(\theta_i^* || x)$$

so that  $C_i(\beta, \gamma_i) = \min_{x \in \mathbb{R}} f(\psi_i, x)$ . Since KL divergences are non-negative,  $f(\psi_i, x)$  is weakly increasing in  $\psi_i$ . To establish the claim, fix two nonnegative numbers  $\psi' < \psi''$ . Let  $x' = \arg \min_x f(\psi', x)$  and  $x'' = \arg \min_x f(\psi'', x)$ . By part (c), these are unique and  $x' < x''$ . Then

$$f(\psi', x') < f(\psi', x'') \leq f(\psi'', x'')$$

where the first inequality uses that  $x' \neq x''$  and  $x'$  is a unique minimum and the second uses the  $f$  is non-decreasing.  $\square$

## F.2 Proof of Proposition 6

We will begin by restating Proposition 6.

**Proposition 6.** *The solution to the optimization problem (12) is the unique allocation  $\psi^*$  satisfying  $\psi_{I^*}^* = \beta$  and*

$$C_i(\beta, \psi_i) = C_j(\beta, \psi_j) \quad \forall i, j \neq I^*. \tag{27}$$

If  $\psi_n = \psi^*$  for all  $n$ , then

$$\Pi_n(\Theta_{I^*}^c) \doteq \exp\{-n\Gamma_\beta^*\}.$$

Moreover under any other adaptive allocation rule, if  $\bar{\psi}_{n,I^*} \rightarrow \beta$  as  $n \rightarrow \infty$  then

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \Pi_n(\Theta_{I^*}^c) \leq \Gamma_\beta^*$$

almost surely.

*Proof.* By Lemma 1, each function  $C_i$  is continuous, and therefore  $\min_{i \neq I^*} C_i(\beta, \psi_i)$  is continuous in  $(\psi_i : i \neq I^*)$ . Since continuous functions on a compact space attain their minimum, there exists an optimal solution  $\psi^*$  to (12), which satisfies

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

Suppose  $\psi^*$  does not satisfy (27), so for some  $j \neq I^*$ ,

$$C_j(\beta, \psi_j^*) > \min_{i \neq I^*} C_i(\beta, \psi_i^*).$$

This yields a contradiction. Consider a new vector  $\psi^\epsilon$  with  $\psi_j^\epsilon = \psi_j^* - \epsilon$  and  $\psi_i^\epsilon = \psi_i^* + \epsilon/(k-2)$  for each  $i \notin \{I^*, j\}$ . For sufficiently small  $\epsilon$ , one has

$$C_j(\beta, \psi_j^\epsilon) > \min_{i \neq I^*} C_i(\beta, \psi_i^\epsilon) > \min_{i \neq I^*} C_i(\beta, \psi_i^*)$$

and so  $\psi^\epsilon$  attains a higher objective value. To show the solution to (27) must be unique, imagine  $\psi$  and  $\psi'$  both satisfy (27) and  $\psi_{I^*} = \psi'_{I^*} = \beta$ . If  $\psi_j > \psi'_j$  for some  $j$ , then  $C_j(\beta, \psi_j) > C_j(\beta, \psi'_j)$  since  $C_j$  is strictly increasing. But by (27) this implies that  $C_j(\beta, \psi_j) > C_j(\beta, \psi'_j)$  for every  $j \neq I^*$ , which implies  $\psi_j > \psi'_j$  for every  $j$ , and contradicts that  $\sum_{j \neq I^*} \psi_j = \sum_{j \neq I^*} \psi'_j = 1 - \beta$ .

The remaining claims follow immediately from Proposition 4 and Lemma 1, which together show that under any adaptive allocation rule

$$\Pi_n(\Theta_{I^*}^c) \doteq \exp\{-n \min_{i \neq I^*} C_i(\bar{\psi}_{n,I^*}, \bar{\psi}_{n,i})\}.$$

This implies that if  $\bar{\psi}_n = \psi^*$  for all  $n$ , then  $\Pi_n(\Theta_{I^*}^c) \doteq \exp\{-n\Gamma_\beta^*\}$ . Similarly, by the continuity of each  $C_i$ , if  $\bar{\psi}_{n,I^*} \rightarrow \beta$ , then

$$\Pi_n(\Theta_{I^*}^c) \doteq \exp\{-n \min_{i \neq I^*} C_i(\beta, \bar{\psi}_{n,i})\} \geq \exp\{-n\Gamma_\beta^*\}$$

which establishes the final claim. □

### F.3 Proof of Lemma 2

Recall, the notation

$$\Gamma^* = \max_{\psi} \min_{i \neq I^*} C_i(\psi_{I^*}, \psi_i) \quad \Gamma_\beta^* \triangleq \max_{\psi: \psi_{I^*} = \beta} \min_{i \neq I^*} C_i(\beta, \psi_i)$$

where

$$C_i(\beta, \psi) = \min_{x \in \mathbb{R}} \beta d(\theta_{I^*}^* || x) + \psi d(\theta_i^* || x).$$



**Lemma 2.** 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  $\Gamma^* \leq 2\Gamma_{1/2}^*$

*Proof.* Define for each non-negative vector  $\boldsymbol{\psi}$ ,

$$f(\boldsymbol{\psi}) = \min_{i \neq I^*} C_i(\psi_{I^*}, \psi_i)$$

The optimal exponent  $\Gamma^*$  is the maximum of  $f(\boldsymbol{\psi})$  over probability vectors  $\boldsymbol{\psi}$ . Here, we instead define  $f$  for all non-negative vectors, and proceed by varying the total budget of measurement effort available  $\sum_{i=1}^k \psi_i$ .

Because each  $C_i$  is non-decreasing (see Lemma 1),  $f$  is non-decreasing. Since the minimum over  $x$  in the definition of  $C_i$  only depends on the relative size of the components of  $\boldsymbol{\psi}$ ,  $f$  is homogenous of degree 1. That is  $f(c\boldsymbol{\psi}) = cf(\boldsymbol{\psi})$  for all  $c \geq 1$ . For each  $c_1, c_2 > 0$  define

$$g(c_1, c_2) = \max \{ f(\boldsymbol{\psi}) : \psi_{I^*} = c_1, \sum_{i \neq I^*} \psi_i \leq c_2, \boldsymbol{\psi} \geq 0 \}.$$

The function  $g$  inherits key properties of  $f$ ; it is also non-decreasing and homogenous of degree 1. We have

$$\begin{aligned} \Gamma_{\beta}^* &= \max \{ f(\boldsymbol{\psi}) : \psi_{I^*} = \beta, \sum_{i=1}^k \psi_i = 1, \boldsymbol{\psi} \geq 0 \} \\ &= \max \{ f(\boldsymbol{\psi}) : \psi_{I^*} = \beta, \sum_{i \neq I^*} \psi_i \leq 1 - \beta, \boldsymbol{\psi} \geq 0 \} \\ &= g(\beta, 1 - \beta) \end{aligned}$$

where the second equality uses that  $f$  is non-decreasing. Similarly,  $\Gamma^* = g(\beta^*, 1 - \beta^*)$ . Setting

$$r := \max \left\{ \frac{\beta^*}{\beta}, \frac{1 - \beta^*}{1 - \beta} \right\}$$

implies  $r\beta \geq \beta^*$  and  $r(1 - \beta) \geq 1 - \beta^*$ . Therefore

$$r\Gamma_{\beta}^* = rg(\beta, 1 - \beta) = g(r\beta, r(1 - \beta)) \geq g(\beta^*, 1 - \beta^*) = \Gamma^*.$$

□

#### F.4 Sub-Gaussian Bound: Proof of Proposition 1

The proof of Proposition 1 relies on the following variational form of Kullback–Leibler divergence, which is given in Theorem 5.2.1 of Robert Gray’s textbook *Entropy and Information Theory* Gray [2011].

**Fact 2.** Fix two probability measures  $\mathbf{P}$  and  $\mathbf{Q}$  defined on a common measurable space  $(\Omega, \mathcal{F})$ . Suppose that  $\mathbf{P}$  is absolutely continuous with respect to  $\mathbf{Q}$ . Then

$$D(\mathbf{P}||\mathbf{Q}) = \sup_X \left\{ \mathbf{E}_{\mathbf{P}}[X] - \log \mathbf{E}_{\mathbf{Q}}[e^X] \right\},$$

where the supremum is taken over all random variables  $X$  such that the expectation of  $X$  under  $\mathbf{P}$  is well defined, and  $e^X$  is integrable under  $\mathbf{Q}$ .

When comparing two normal distributions  $\mathcal{N}(\theta, \sigma^2)$  and  $\mathcal{N}(\theta', \sigma^2)$  with common variance, the KL-divergence can be expressed as  $d(\theta||\theta') = (\theta - \theta')^2/(2\sigma^2)$ . We follow [Russo and Zou \[2015\]](#) in deriving the following corollary of [Fact 2](#), which provides and analogous lower bound on the KL-divergences when distributions are sub-Gaussian. Recall that,  $\mu(\theta) = \int yp(y|\theta)d\nu(y)$  denotes the mean observation under  $\theta$ .

**Corollary 3.** *Fix any  $\theta, \theta' \in [\underline{\theta}, \bar{\theta}]$ . If when  $Y \sim p(y|\theta')$ ,  $Y$  is sub-Gaussian with parameter  $\sigma$ , then,*

$$d(\theta||\theta') \geq \frac{(\mu(\theta) - \mu(\theta'))^2}{2\sigma^2}$$

*Proof.* Consider two alternate probability distributions for a random variable  $Y$ , one where  $Y \sim p(y|\theta)$  and one where  $Y \sim p(y|\theta')$ . We apply [Fact 2](#) where  $X = \lambda(Y - \mathbf{E}_{\theta'}[Y])$ ,  $\mathbf{P}$  is the probability measure when  $Y \sim p(y|\theta)$  and  $\mathbf{Q}$  is the measure when  $Y \sim p(y|\theta')$ . By the sub-Gaussian assumption  $\log \mathbf{E}_{\theta'}[\exp\{X\}] \leq \lambda^2\sigma^2/2$ . Therefore, [Fact 2](#) implies

$$d(\theta||\theta') \geq \lambda(\mathbf{E}_{\theta}[X]) - \frac{\lambda^2\sigma^2}{2} = \lambda(\mathbf{E}_{\theta}[Y] - \mathbf{E}_{\theta'}[Y]) - \frac{\lambda^2\sigma^2}{2}.$$

The result follows by choosing  $\lambda = (\mathbf{E}_{\theta}[Y] - \mathbf{E}_{\theta'}[Y])/\sigma^2$  which minimizes the right hand side.  $\square$

We are now ready to prove [Proposition 1](#). Recall that in an exponential family,  $A'(\theta) = \int T(y)p(y|\theta)d\nu(y)$ , so if  $T(y) = y$  then  $A'(\theta) = \mu(\theta)$ .

*Proof of Proposition 1.* By [Lemma 1](#),

$$\Gamma_{1/2}^* = \max_{\psi: \psi_{I^*} = 1/2} \min_{i \neq I^*} C_i(1/2, \psi_i)$$

Let  $\mu_{I^*} = A'(\theta_{I^*}^*)$  and  $\mu_i = A'(\theta_i^*)$  denote the means of designs  $I^*$  and  $i$  so  $\Delta_i = \mu_{I^*} - \mu_i$ . By [Lemma 1](#),

$$C_i(1/2, \psi_i) = (1/2)d(\theta_{I^*}^*||\bar{\theta}) + \psi_i d(\theta_i^*||\bar{\theta}).$$

where  $\bar{\theta}$  is the unique parameter with mean

$$A'(\bar{\theta}) = \frac{(1/2)\mu_{I^*} + \psi_i\mu_i}{1/2 + \psi_i}.$$

For  $\psi_i \leq 1/2$ ,

$$A'(\bar{\theta}) \geq \frac{\mu_{I^*} + \mu_i}{2} = \mu_i + \Delta_i/2.$$

Now, using [Corollary 3](#) and the non-negativity of KL-divergence

$$C_i(1/2, \psi_i) \geq \psi_i d(\theta_i^*||\bar{\theta}) \geq \frac{\psi_i(\mu_i - \mu_i + \Delta_i/2)^2}{2\sigma^2} = \frac{\psi_i\Delta_i^2}{8\sigma^2}.$$

Choosing  $\psi_{I^*} = 1/2$ , and  $\psi_i \propto \Delta_i^{-2}$ , so

$$\psi_i = \frac{1}{2} \left( \sum_{j=2}^k \Delta_j^{-2} \right)^{-1} \Delta_i^{-2}$$

yields

$$\min_{i \neq I^*} C_i(1/2, \psi_i) \geq \frac{1}{16\sigma^2 \sum_2^k \Delta_j^{-2}}.$$

$\square$

## F.5 Convergence of Uniform Allocation: Proof of Proposition 2

*Proof.* Without loss of generality, assume the problem is parameterized so that the mean of design  $i$  is  $\theta_i^*$ . By Proposition 5, we have

$$\Pi_n(\Theta_{I^*}^c) \doteq \exp\{-n \min_{i \neq I^*} C_i(k^{-1}, k^{-1})\}$$

By Lemma 1,

$$C_i(k^{-1}, k^{-1}) = k^{-1}d(\theta_{I^*}^* || \bar{\theta}) + k^{-1}d(\theta_i^* || \bar{\theta})$$

where  $\bar{\theta} = (\theta_{I^*}^* + \theta_i^*)/2$ . Therefore, using the formula for the KL-divergence of standard Gaussian random variables

$$C_i(k^{-1}, k^{-1}) = \frac{(\theta_{I^*}^* - \bar{\theta})^2}{2\sigma^2} + \frac{(\theta_i^* - \bar{\theta})^2}{2\sigma^2} = \frac{(\theta_{I^*}^* - \theta_i^*)^2}{4\sigma^2} = \frac{\Delta_i^2}{4\sigma^2}.$$

□

## G Analysis of the Top-Two Allocation Rules: Proof of Proposition 7

**Proposition 7.** *Under the TTTS, TTPS, or TTVS algorithm with parameter  $\beta > 0$ ,  $\bar{\psi}_n \rightarrow \psi^\beta$ , where  $\psi^\beta$  is the unique allocation with  $\psi_{I^*}^\beta = \beta$  satisfying*

$$C_i(\beta, \psi_i^\beta) = C_j(\beta, \psi_j^\beta) \quad \forall i, j \neq I^*. \quad (28)$$

Therefore,

$$\Pi_n(\Theta_{I^*}^c) \doteq e^{-n\Gamma_\beta^*}. \quad (29)$$

Because each  $C_i$  is continuous, if  $\bar{\psi}_n \rightarrow \psi^\beta$  then  $C_i(\bar{\psi}_{n,I^*}, \bar{\psi}_{n,i}) \rightarrow C_i(\beta, \psi_i^\beta)$  for all  $i \neq I^*$ . Equation (29) then follows by invoking Proposition 6, which establishes the optimality of the allocation  $\psi^\beta$ .

The remainder of this section establishes that  $\bar{\psi}_n \rightarrow \psi^\beta$  almost surely the proposed top-two rules. The proof is broken into a number of steps. In order to provide a nearly unified treatment of the three algorithms, we begin with several results that hold for any allocation rule.

### G.1 Results for a general allocation rule

As in other sections, all arguments here hold for any sample path (up to a set of measure zero). The first result provides a sufficient condition under which  $\bar{\psi}_n \rightarrow \psi^\beta$ . Roughly speaking, if  $\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta$ , then too much measurement effort has been allocated to design  $j$  relative to the optimal proportion  $\psi_j^\beta$ . Algorithms satisfying (30) allocate negligible measurement effort to such designs, and therefore the average measurement effort they receive must decrease toward the optimal proportion.

**Lemma 11** (Sufficient condition for optimality). *Consider any adaptive allocation rule. If  $\bar{\psi}_{n,I^*} \rightarrow \beta$  and*

$$\sum_{n \in \mathbb{N}} \psi_{n,j} \mathbf{1}(\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta) < \infty \quad \forall j \neq I^*, \delta > 0, \quad (30)$$

then  $\bar{\psi}_n \rightarrow \psi^\beta$ .

*Proof.* Fix a sample path for which  $\psi_{n,I^*} \rightarrow \beta$ , and (30) holds. Fix some  $j \neq I^*$ . We first show  $\liminf_{n \rightarrow \infty} \bar{\psi}_{n,j} \leq \psi_j^*$ . Suppose otherwise. Then, with positive probability, for some  $\delta > 0$ , there exists  $N$  such that for all  $n \geq N$ ,  $\bar{\psi}_{n,j} \geq \psi_j^* + \delta$ . But then,

$$\sum_{n \in \mathbb{N}} \psi_{n,j} = \sum_{n=1}^N \psi_{n,j} + \sum_{n=N+1}^{\infty} \mathbf{1}(\bar{\psi}_{n,j} \geq \psi_j^* + \delta) \psi_{n,j} < \infty.$$

But since  $\bar{\psi}_{n,j} = \sum_{\ell=1}^n \psi_{\ell,j}/n$  this implies  $\bar{\psi}_{n,j} \rightarrow 0$ .

Now, we show  $\limsup_{n \rightarrow \infty} \bar{\psi}_{n,j} \leq \psi_j^*$ . Proceeding by contradiction again, suppose otherwise. Then, with positive probability

$$\limsup_{n \rightarrow \infty} \bar{\psi}_{n,j} > \psi_j^\beta \quad \& \quad \liminf_{n \rightarrow \infty} \bar{\psi}_{n,j} \leq \psi_j^\beta.$$

On any sample path where this occurs, for some  $\delta > 0$ , there exists an infinite sequence of times  $N_1 < N_2 < N_3 < \dots$  such that  $\bar{\psi}_{N_\ell,j} \geq \psi_j^\beta + 2\delta$  when  $\ell$  is odd and  $\bar{\psi}_{N_\ell,j} \leq \psi_j^\beta + \delta$  when  $\ell$  is even. This can only occur if,

$$\sum_{n \in \mathbb{N}} \psi_{n,j} \mathbf{1}(\bar{\psi}_{n,j} \geq \psi_j^* + \delta) = \infty,$$

which violates the hypothesis.

Together with the hypothesis that  $\bar{\psi}_{n,I^*} \rightarrow \beta$ , this implies that for all  $i \in \{1, \dots, k\}$ ,  $\limsup_{n \rightarrow \infty} \bar{\psi}_{n,i} \leq \psi_i^\beta$ . But since  $\sum_i \bar{\psi}_{n,i} = \sum_i \psi_i^\beta$ , this implies  $\bar{\psi}_n \rightarrow \psi^\beta$ .  $\square$

The next lemma will be used to establish that (30) holds for each of the proposed algorithms. It shows that if too much measurement effort has been allocated to some design  $i \neq I^*$ , in the sense that  $\bar{\psi}_{n,i} > \psi_i^\beta + \delta$  for a constant  $\delta > 0$ , then  $\alpha_{n,i}$  is exponentially small compared  $\max_{j \neq I^*} \alpha_{n,j}$ .

**Lemma 12** (Over-allocation implies negligible probability). *Fix any  $\delta > 0$  and  $j \neq I^*$ . With probability 1, under any allocation rule, if  $\bar{\psi}_{n,I^*} \rightarrow \beta$ , there exists  $\delta' > 0$  and a sequence  $\epsilon_n$  with  $\epsilon_n \rightarrow 0$  such that for any  $n \in \mathbb{N}$ ,*

$$\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta \implies \frac{\alpha_{n,j}}{\max_{i \neq I^*} \alpha_{n,i}} \leq e^{-n(\delta' + \epsilon_n)}.$$

*Proof.* Since  $\Pi_n(\Theta_{I^*}^c) = \sum_{i \neq I^*} \alpha_{n,i}$ ,  $\Pi_n(\Theta_{I^*}^c) \doteq \max_{i \neq I^*} \alpha_{n,i}$ . Then, by invoking Proposition (6), since  $\bar{\psi}_{n,I^*} \rightarrow \beta$ ,

$$\limsup_{n \rightarrow \infty} -\frac{1}{n} \log \left( \max_{i \neq I^*} \alpha_{n,i} \right) \leq \Gamma_\beta^*.$$

Recall the definition  $\bar{\Theta}_i \triangleq \{\theta | \theta_i \geq \theta_{I^*}\}$ . Now, by Proposition 4 and Lemma 1,

$$\alpha_{n,j} = \Pi_n(\Theta_j) \leq \Pi_n(\bar{\Theta}_j) \doteq \exp\{-nC_j(\bar{\psi}_{n,I^*}, \bar{\psi}_{n,j})\} \doteq \exp\{-nC_j(\beta, \bar{\psi}_{n,j})\}.$$

Combining these equations implies that there exists a non-negative sequence  $\epsilon_n \rightarrow 0$  with

$$\frac{\alpha_{n,j}}{\max_{i \neq I^*} \alpha_{n,i}} \leq \frac{\exp\{-n(C_j(\beta, \bar{\psi}_{n,j}) - \epsilon_n/2)\}}{\exp\{-n(\Gamma_\beta^* + \epsilon_n/2)\}} = \exp\left\{-n \left( (C_j(\beta, \bar{\psi}_{n,j}) - \Gamma_\beta^*) - \epsilon_n \right)\right\}$$

Since  $C_j(\beta, \psi_j)$  is strictly increasing in  $\psi_j$  (See lemma 1) and  $C_j(\beta, \psi_j^\beta) = \Gamma_\beta^*$ , there exists some  $\delta' > 0$  such that

$$\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta \implies C_j(\beta, \bar{\psi}_{n,j}) - \Gamma_\beta^* > \delta'.$$

$\square$

The next result builds on Proposition 3. It shows that the quality of any design which receives infinite measurement effort is identified to arbitrary precision. On the other hand, for designs receiving finite measurement effort, there is always nonzero probability under the posterior that one of them significantly exceeds the highest quality that has been confidently identified. Therefore,  $\alpha_{n,i}$  and  $V_{n,i}$  remain bounded away from 0 for designs that receive finite measurement effort. This result will be used to show that all designs receive infinite measurement effort under the proposed top-two allocation rules, and as a result the posterior converges on the truth asymptotically.

**Lemma 13** (Implications of finite measurement). *Let*

$$\mathcal{I} = \{i \in \{1, \dots, k\} : \sum_{n=1}^{\infty} \psi_{n,i} < \infty\}$$

denote the set of designs to which a finite amount of measurement effort is allocated. Then, for any  $i \notin \mathcal{I}$

$$\Pi_n(\{\boldsymbol{\theta} : \theta_i \in (\theta_i^* - \epsilon, \theta_i^* + \epsilon)\}) \rightarrow 1, \quad (31)$$

and if  $\mathcal{I}$  is empty

$$V_{n,i} \rightarrow \begin{cases} 0 & \text{if } i \neq I^* \\ v_{I^*}(\boldsymbol{\theta}^*) > 0 & \text{if } i = I^* \end{cases} \quad \text{and} \quad \alpha_{n,i} \rightarrow \begin{cases} 0 & \text{if } i \neq I^* \\ 1 & \text{if } i = I^*. \end{cases}$$

If  $\mathcal{I}$  is nonempty, then for every  $i \in \mathcal{I}$ ,

$$\liminf_{n \rightarrow \infty} \alpha_{n,i} > 0 \quad \text{and} \quad \liminf_{n \rightarrow \infty} V_{n,i} > 0.$$

*Proof.* Equation (31) is implied by Proposition 3. Now, set

$$\Theta_{i,\epsilon} = \{\boldsymbol{\theta} \in \Theta : \theta_i \geq \max_{j \neq i} \theta_j + \epsilon\}$$

to be the set of parameters under which the quality of design  $i$  exceeds that of all others by at least  $\epsilon$ . Let  $\rho^* = \max_{i \notin \mathcal{I}} \theta_i^*$  denote the quality of the best design among those that are sampled infinitely often, and choose  $\epsilon > 0$  small enough that  $\rho^* + 2\epsilon < \bar{\theta}$ . For  $i \in \mathcal{I}$ , we have

$$\Pi_n(\Theta_{i,\epsilon}) \geq \Pi_n(A) - \Pi_n(B)$$

for

$$A \equiv \{\boldsymbol{\theta} | \theta_i \geq \rho^* + 2\epsilon \ \& \ \theta_j < \rho^* \ \forall j \in \mathcal{I} \setminus \{i\}\}$$

defined to be parameters under which  $\theta_i \geq \rho^* + 2\epsilon$  but none of the other designs in  $\mathcal{I}$  exceed  $\rho^*$ , and

$$B \equiv \{\boldsymbol{\theta} : \max_{i \notin \mathcal{I}} \theta_i \geq \rho^* + \epsilon\}$$

defined to be the parameter vectors under which there is no design in  $\mathcal{I}^c$  with quality exceeding  $\rho^* + \epsilon$ . By (31),

$$\Pi_n(B) \rightarrow 0,$$

but by the second part of Proposition 3, the set of parameters  $A$  cannot be completely ruled based on a finite amount of measurement effort, and

$$\inf_{n \in \mathbb{N}} \Pi_n(A) > 0.$$

Together this shows

$$\liminf_{n \rightarrow \infty} \Pi_n(\Theta_{i,\epsilon}) > 0,$$

which implies the result. □

## G.2 Results specific to the proposed algorithms

We now leverage the general results of the previous subsection to show  $\bar{\psi} \rightarrow \psi^\beta$  under each proposed top-two allocation rule. Proofs are provided separately for each of the three algorithms, but they follow a similar structure. In the first step, we use Lemma 13 to argue that  $\bar{\psi}_{n,I^*} \rightarrow \beta$  almost surely. The proof then uses Lemma 12 to show (30) holds, which by Lemma 11 is sufficient to establish that  $\bar{\psi}_n \rightarrow \psi^\beta$ .

### G.2.1 Top-Two Thompson Sampling

Recall that under top-two Thompson sampling, for every  $i \in \{1, \dots, k\}$ ,

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

*Proof for TTTS.*

*Step 1: Show  $\bar{\psi}_{n,I^*} \rightarrow \beta$ .* To begin, we show  $\sum_{n \in \mathbb{N}} \psi_{n,i} = \infty$  for each design  $i$ . Suppose otherwise. Let  $\mathcal{I} = \{i \in \{1, \dots, k\} : \sum_1^\infty \psi_{n,i} < \infty\}$  be the set of designs to which finite measurement effort is allocated. Under the TTTS sampling rule,  $\psi_{n,i} \geq \beta \alpha_{n,i}$ . Therefore, by Lemma 13, if  $i \in \mathcal{I}$  then  $\liminf_{n \rightarrow \infty} \alpha_{n,i} > 0$ , which implies  $\sum_{n \in \mathbb{N}} \psi_{n,i} = \infty$ , a contradiction.

Since  $\sum_1^\infty \psi_{n,i} = \infty$  for all  $i$ , by applying Lemma 13 we conclude that  $\alpha_{n,I^*} \rightarrow 1$ . For TTTS, this implies  $\bar{\psi}_{n,I^*} \rightarrow \beta$ .

*Step 2: Show (30) holds.* By Lemma 11, it is enough to show that (30) holds under TTTS. Let  $\hat{I}_n = \arg \max_i \alpha_{n,i}$ , and  $\hat{J}_n = \arg \max_{i \neq \hat{I}_n} \alpha_{n,i}$ . Since  $\alpha_{n,I^*} \rightarrow 1$ , for each sample path there is a finite time  $\tau < \infty$  such that for all  $n \geq \tau$ ,  $\hat{I}_n = I^*$  and therefore  $\hat{J}_n = \arg \max_{i \neq I^*} \alpha_{n,i}$ . Under TTTS,

$$\psi_{n,i} \leq \beta \alpha_{n,i} + (1 - \beta) \frac{\alpha_{n,i}}{\alpha_{n,\hat{J}_n}} \leq \frac{\alpha_{n,i}}{\alpha_{n,\hat{J}_n}},$$

where the first inequality follows since

$$\sum_{j \neq i} \frac{\alpha_{n,j}}{1 - \alpha_{n,j}} \leq \frac{\sum_{j \neq i} \alpha_{n,i}}{1 - \alpha_{n,\hat{I}_n}} \leq \frac{\sum_{j \neq i} \alpha_{n,j}}{\alpha_{n,\hat{J}_n}} \leq \frac{1}{\alpha_{n,\hat{J}_n}}.$$

For  $n \geq \tau$ , this means  $\psi_{n,i} \leq \alpha_{n,i} / (\max_{j \neq I^*} \alpha_{n,i})$  for any  $i \neq I^*$ . By Lemma 12, there is a constant  $\delta' > 0$  and a sequence  $\epsilon_n \rightarrow 0$  such that

$$\bar{\psi}_{n,i} \geq \psi_i^\beta + \delta \implies \frac{\alpha_{n,i}}{\max_{j \neq I^*} \alpha_{n,j}} \leq e^{-n(\delta' - \epsilon_n)}.$$

Therefore for all  $i \neq I^*$

$$\sum_{n \geq \tau} \psi_{n,i} \mathbf{1}(\bar{\psi}_{n,i} \geq \psi_i^\beta + \delta) \leq \sum_{n \geq \tau} e^{-n(\delta' - \epsilon_n)} < \infty.$$

□

## G.2.2 Top-Two Probability Sampling

Recall that top-two probability sampling sets  $\psi_{n,\hat{I}_n} = \beta$  and  $\psi_{n,\hat{J}_n} = 1 - \beta$  where  $\hat{I}_n = \arg \max_i \alpha_{n,i}$  and  $\hat{J}_n = \arg \max_{j \neq \hat{I}_n} \alpha_{n,i}$  are the two designs with the highest posterior probability of being optimal.

*Proof for TTPS.*

*Step 1: Show  $\bar{\psi}_{n,I^*} \rightarrow \beta$ .* To begin, we show  $\sum_{n \in \mathbb{N}} \psi_{n,i} = \infty$  for each design  $i$ . Suppose otherwise. Let  $\mathcal{I} = \{i \in \{1, \dots, k\} : \sum_1^\infty \psi_{n,i} < \infty\}$  be the set of designs to which finite measurement effort is allocated. Proceeding by contradiction, suppose  $\mathcal{I}$  is nonempty. By Lemma 13, there is a time  $\tau$  and some probability  $\alpha' > 0$  such that  $\alpha_{n,i} > \alpha'$  for all  $n \geq \tau$  and  $i \in \mathcal{I}$ . However, because of the assumption that  $\theta_i^* \neq \theta_j^*$ , for  $i \neq j$ ,  $I = \arg \max_{i \notin \mathcal{I}} \theta_i^*$  is unique. By (31), the algorithm identifies  $\arg \max_{i \notin \mathcal{I}} \theta_i^*$  with certainty, and  $\alpha_{n,i} \rightarrow 0$  for every  $i \notin \mathcal{I}$  except for  $I$ . This means there is a time  $\tau' > \tau$  such that for  $n \geq \tau'$

$$\begin{aligned} \alpha_{n,i} &> \alpha' && \text{if } i \in \mathcal{I} \\ \alpha_{n,i} &\leq \alpha' && \text{if } i \notin \mathcal{I} \text{ and } i \neq I. \end{aligned}$$

When this occurs at least one of the two designs with highest probability  $\alpha_{n,i}$  of being optimal must be in the set  $\mathcal{I}$ , which implies designs in  $\mathcal{I}$  receive infinite measurement effort, yielding a contradiction.

Since  $\sum_1^\infty \psi_{n,i} = \infty$  for all  $i$ , Lemma 13 implies  $\alpha_{n,I^*} \rightarrow 1$ . Therefore, there is a finite time  $\tau$  such that  $\hat{I}_n \triangleq \arg \max_i \alpha_{n,i} = I^*$  for all  $n \geq \tau$ . By the definition of the algorithm  $\psi_{n,\hat{I}_n} = \beta$ , and so  $\psi_{n,I^*} = \beta$  for all  $n \geq \tau$ . We conclude that  $\bar{\psi}_{n,I^*} \rightarrow \beta$ .

*Step 2: Show (30) holds.* As argued above, for each sample path there is a finite time  $\tau < \infty$  such that for all  $n \geq \tau$ ,  $\hat{I}_n = I^*$  and therefore  $\hat{J}_n = \arg \max_{i \neq I^*} \alpha_{n,i}$ . By Lemma 12, one can choose  $\tau' \geq \tau$  such that for all  $n \geq \tau'$ ,

$$\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta \implies \alpha_{n,j} < \max_{i \neq I^*} \alpha_{n,i}$$

and therefore by definition  $\hat{J}_n \neq j$ . This concludes the proof, as it shows that for each sample path there is a finite time  $\tau'$  after which TTPS never allocates any measurement effort to design  $j$  when  $\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta$ .  $\square$

## G.2.3 Top-Two Value Sampling

Recall that top-two value sampling sets and  $\psi_{n,\hat{I}_n} = \beta$  and  $\psi_{n,\hat{J}_n} = 1 - \beta$  where  $\hat{I}_n = \arg \max_i V_{n,i}$  and  $\hat{J}_n = \arg \max_{j \neq \hat{I}_n} V_{n,i}$  are the two designs with the highest posterior value.

*Proof for TTVS. Step 1: Show  $\bar{\psi}_{n,I^*} \rightarrow \beta$ .* The proof is essentially identical to that for TTPS. To begin, we show  $\sum_{n \in \mathbb{N}} \psi_{n,i} = \infty$  for each design  $i$ . Suppose otherwise. Let  $\mathcal{I} = \{i \in \{1, \dots, k\} : \sum_1^\infty \psi_{n,i} < \infty\}$  be the set of designs to which finite measurement effort is allocated. Proceeding by contradiction, suppose  $\mathcal{I}$  is nonempty. By Lemma 13, there is a time  $\tau$  and some  $v > 0$  such that  $V_{n,i} > v$  for all  $n \geq \tau$  and  $i \in \mathcal{I}$ . However, because of the assumption that  $\theta_i^* \neq \theta_j^*$ , for  $i \neq j$ ,  $I = \arg \max_{i \notin \mathcal{I}} \theta_i^*$  is unique<sup>3</sup>. By (31), the algorithm identifies  $\arg \max_{i \notin \mathcal{I}} \theta_i^*$  with certainty, and

<sup>3</sup>If the arg-max is not unique, then one could show that  $V_{n,i} \rightarrow 0$  for all  $i \notin \mathcal{I}$ , and that therefore there is a finite time after both of the top-two designs are always in the set  $\mathcal{I}$ , yielding a contradiction

$V_{n,i} \rightarrow 0$  for every  $i \notin \mathcal{I}$  except for  $I$ . Then there is a time  $\tau' > \tau$  such that for  $n \geq \tau'$

$$\begin{aligned} V_{n,i} &> v && \text{if } i \in \mathcal{I} \\ V_{n,i} &\leq v && \text{if } i \notin \mathcal{I} \text{ and } i \neq I^*. \end{aligned}$$

When this occurs at least one of the two designs with highest value  $V_{n,i}$  must be in the set  $\mathcal{I}$ , which implies designs in  $\mathcal{I}$  receive infinite measurement effort, yielding a contradiction.

Since  $\sum_1^\infty \psi_{n,i} = \infty$  for all  $i$ , Lemma 13 implies  $V_{n,I^*} \rightarrow v_{I^*}(\boldsymbol{\theta}^*) > 0$  and  $V_{n,i} \rightarrow 0$  for all  $i \neq I^*$ . Therefore, there is a finite time  $\tau$  such that  $\arg \max_i V_{n,i} = I^*$  for all  $n \geq \tau$ . By the definition of the algorithm  $\arg \max_i V_{n,i}$  is sampled with probability  $\beta$ , and so  $\psi_{n,I^*} = \beta$  for all  $n \geq \tau$ . We conclude that  $\bar{\psi}_{n,I^*} \rightarrow \beta$ .

*Step 2: Show (30) holds.* Again, the proof is essentially identical to that for TTPS. As argued above, for each sample path there is a finite time  $\tau < \infty$  such that for all  $n \geq \tau$ ,  $\hat{I}_n = I^*$  and therefore  $\hat{J}_n = \arg \max_{i \neq I^*} V_{n,i}$ . By Lemma 3,  $V_{n,i} \doteq \alpha_{n,i}$ . Combining this with Lemma 12 shows one can choose  $\tau' \geq \tau$  such that for all  $n \geq \tau'$ ,

$$\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta \implies V_{n,j} < \max_{i \neq I^*} V_{n,i}$$

and therefore by definition  $\hat{J}_n \neq j$ . This concludes the proof, as it shows that for each sample path there is a finite time  $\tau'$  after which TTVS never allocates any measurement effort to design  $j \neq I^*$  when  $\bar{\psi}_{n,j} \geq \psi_j^\beta + \delta$ .  $\square$