Simple Bayesian Algorithms for Best Arm Identification

Daniel Russo

Microsoft Research New England, djrusso@Microsoft.com

February 29, 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. I propose three simple Bayesian algorithms for adaptively allocating measurement effort. One 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 a 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. I prove that these simple algorithms satisfy a strong optimality property. In a frequestist setting where the true quality of the designs is fixed, one hopes the posterior definitively identifies the optimal design, in the sense that that the posterior probability assigned to the event that some other design is optimal converges to zero as measurements are collected. I 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 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. Problems of this form have been studied heavily in operations research, statistics, and computer science, as they form a useful abstraction of 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?

This paper proposes three simple algorithms for allocating measurement effort. Each algorithm begins with a prior distribution over the unknown quality of the designs. As measurements are gathered, the experimenter learns about the designs, and beliefs are updated to form a posterior distribution. This posterior distribution gives a principled method for reasoning about the uncertain quality of different designs, and for assessing the probability any given design is optimal.

The first algorithm I 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 probability $\beta > 0$, 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. The *Top–Two Value sampling* algorithm proceeds in a similar manner, but in estimating the top-two designs it considers not only the probability a design is optimal, but the expected amount by which it exceeds other designs.

The final algorithm I propose is a modification 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 [2014], 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 [2014a]), but it's designed to maximize the cumulative reward earned while sampling. As a result, in the long run the algorithm allocates almost all effort to measuring the estimated-best design, and it takes a long time to certify that none of the alternative designs would offer better performance. This paper shows, however, that the algorithm can be easily modified for the objective of efficiently identifying the best arm. The variant I propose adds a simple re-sampling step to Thompson sampling, which prevents the algorithm from allocating too much measurement effort to the estimated-best design.

These simple heuristic algorithms are shown to satisfy a strong optimality property. The 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. I show that under the proposed algorithms this convergence occurs at an *exponential rate*, and the corresponding exponent is essentially the best possible among all allocation rules.

In studying the optimality of the proposed algorithms, the paper characterizes what the simulation optimization literature calls an optimal computing budget allocation (Chen et al. [2000]). This specifies the asymptotically optimal fraction of measurement effort to allocate to each design as a complicated function of the true quality of the designs. Rather than using an *equal allocation* of measurement effort across designs, this calculation shows it is optimal to adjust measurement effort to gather *equal evidence* to rule out each sub-optimal design. By only sampling designs that seem most promising under the posterior, each of the proposed algorithms automatically balances the evidence gathered in this manner.

While the paper's main theoretical results are asymptotic, this is mainly to provide sharp insight. The algorithm design is completely separate from the theoretical analysis, so any approximations in the analysis don't influence practical performance. Numerical experiments suggest the proposed algorithms perform very well over moderate horizons.

2 Related Literature

Thompson sampling, and other simple Bayesian algorithms have attracted a great deal of recent interest in the multi-armed bandit literature. These algorithms have been studied for classical bandit problems (Agrawal and Goyal [2012], Korda et al. [2013], Kaufmann et al. [2012]), as well as much more complicated types of online optimization problems (Srinivas et al. [2012], Gopalan et al. [2014], Russo and Van Roy [2014a,b], Johnson et al. [2015]). This paper shows how to extend some of these successful techniques to a pure-exploration problem, where the goal is quickly to identify a near-optimal action. I hope the new techniques here will aid in the development and analysis of adaptive Bayesian algorithms for other problems.

Many machine learning papers have studied the problem of identifying the best arm in a multi-armed bandit problem. (Even-Dar et al. [2002], Mannor and Tsitsiklis [2004], Audibert and Bubeck [2010], Gabillon et al. [2012], Karnin et al. [2013], Kaufmann and Kalyanakrishnan [2013], Jamieson and Nowak [2014], Kaufmann et al. [2014]) In the PAC-learning framework, Even-Dar et al. [2002] shows that when measurement noise is uniformly bounded, an ϵ–optimal arm can be identified with probability at least δ using $O\left((k/\epsilon^2)\log(1/\delta)\right)$ samples. Mannor and Tsitsiklis [2004] and Even-Dar et al. [2006] provide matching lower bounds, more refined stopping criteria, and distribution specific results, which bound the number of samples required as a function of the (unknown) quality of the deigns. Audibert and Bubeck [2010] provide algorithms under which the probability of incorrect selection decay exponentially as samples are collected. All of these results assume measurement noise is uniformly bounded, and provide rates of convergence as a function only of the (unknown) means of the distributions. This leaves open the question of exactly characterizing the complexity of identifying the best arm; one would hope for a Lai-Robbins style bound (Lai and Robbins [1985]) that provides exact constants and depends on an information-theoretic measure of the separation between distributions instead of the difference in their means. Significant recent progress along these lines was presented by Kaufmann and Kalyanakrishnan [2013] and Kaufmann et al. [2014], but there are still gaps between the upper and lower bounds. I hope this work contributes to closing this gap, and conjecture that the optimal exponent Γ^* studied in this paper is the right asymptotic complexity measure for the so-called "fixed-confidence" setting. Along these lines, I provide a lower bound for that problem involving Γ^* in the appendix.

This problem has also been studied in the simulation optimization literature, where it is typically called either *ranking and selection*, or *ordinal optimization*. This work is most closely related to the optimal computing budget allocation (OCBA) of Chen et al. [2000]. Assuming measurement noise is Gaussian, the fraction of measurement effort allocated to each design is fixed over time, and the algorithm returns the design with the highest empirical mean, Chen et al. [2000] derives asymptotic approximations to the probability of incorrect selection as a function of the quality of the designs. The OCBA is the fixed allocation that minimizes this asymptotic error approximation. Glynn and Juneja [2004] provide a large-deviations perspective, which gives insight into the structure of the OCBA and allows for generalization beyond Gaussian distributions. The simplified form of the optimal exponent given in Section 7.3, is closely related to the calculations of Glynn and Juneja [2004]. Our paper derives an analogue of the OCBA for our problem as a function of the unknown quality of the designs. Unlike this line of work, however, we show that this allocation is optimal even among adaptive strategies, and focus on establishing that simple adaptive schemes can match its rate of convergence.

Finally, this work was highly influenced by Chernoff's classic paper (Chernoff [1959]) on the sequential design of experiments for binary hypothesis testing. Although Chernoff's paper is rarely mentioned in this literature, the best arm identification problem is closely related to the problem he studies, and his asymptotic derivations give great insight into this problem. The optimal exponent

given in Section 7.3, which governs the rate of convergence of the posterior distribution, closely mirrors the sample complexity measure of Chernoff [1959].

3 Problem Formulation

We 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 $I_n \in \{1, ..., k\}$, and observes a measurement Y_{n,I_n}. The measurement $Y_{n,i} \in \mathbb{R}$ associated with design i and time n is drawn from a fixed, unknown, probability distribution, and the vector $\boldsymbol{Y}_n \triangleq (Y_{n,1}, ..., Y_{n,k})$ 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 sequence of observations $I_1, Y_{1,I_1}, ...I_{n-1}, Y_{n-1,I_{n-1}}$. The goal is to efficiently identify the design with the highest mean.

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

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

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

Throughout the paper, $\boldsymbol{\theta}^* \triangleq (\theta_1^*, ..., \theta_k^*)$ 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. I will assume throughout that $\theta_i^* \neq \theta_j^*$ for $i \neq j$ so that I^* is unique, although this can be relaxed by considering an indifference zone formulation where the goal is to identify an ϵ–optimal design, for some specified tolerance level $\epsilon > 0$.

Prior and Posterior Distributions. The policies studied in this paper make use of a prior distribution Π_0 over a set of possible parameters Θ that contains $\boldsymbol{\theta}^*$. Based on a sequence of observations $(I_1, Y_{1,I_1}, ..., I_{n-1}, Y_{n,I_n})$, beliefs are updated to attain a posterior distribution Π_n. I assume Π_0 has density π_0 with respect to lebesque measure. In this case, the posterior distribution Π_n has corresponding density

$$\pi_n(\boldsymbol{\theta}) = \frac{\pi_0(\boldsymbol{\theta}) L_n(\boldsymbol{\theta})}{\int_\Theta \pi_0(\boldsymbol{\theta}') L_n(\boldsymbol{\theta}') d\boldsymbol{\theta}'}, \tag{1}$$

where

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

is the likelihood function. It's worth highlighting that 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

$$\Theta_i = \left\{ \boldsymbol{\theta} \in \Theta \middle| \theta_i = \max_{j \in \{1,...,k\}} \theta_j \right\}$$

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

$$\alpha_{n,i} \triangleq \Pi_n(\Theta_i) = \int_{\Theta_i} \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2}$$

denote the posterior probability assigned to the event that action i is optimal. Define $\Theta_{-I^*} \triangleq \bigcup_{i \neq I^*} \Theta_i$ to be the set of parameters under which I^* is not optimal. Our analysis will focus on

$$\Pi_n(\Theta_{-I^*}) = 1 - \alpha_{I^*},$$

which is the posterior probability assigned to the event that an action other than I^* is optimal. The next section will introduce policies under which $\Pi_n(\Theta_{-I^*}) \longrightarrow 0$ as $n \longrightarrow \infty$, 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 $(I_1, Y_{1,I_1}, ...I_{n-1}, Y_{n-1,I_{n-1}})$. For all $i \in \{1, ..., k\}$ and $n \in \mathbb{N}$, define

$$\psi_{n,i} \triangleq \mathbf{P}(I_n = i | \mathcal{F}_n) \qquad \overline{\psi}_{n,i} \triangleq \frac{1}{n} \sum_{\ell=1}^{n} \psi_{\ell,i} \qquad S_{n,i} \triangleq \sum_{i=1}^{n} \mathbf{1}(I_n = i).$$

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

4 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 $\hat{I} \in \{1, ..., k\}$ of the optimal design, and measures that with probability β. Otherwise, we consider a counterfactual: in the (possibly unlikely) event that $\hat{I}$ is not the best design, which alternative $\hat{J} \neq \hat{I}$ is most likely to be the best design? With probability $1 - \beta$, the algorithm measures the alternative $\hat{J}$. The algorithms differ in how they compute $\hat{I}$ and $\hat{J}$. The most computationally efficient is the modified version of Thompson sampling, under which $\hat{I}$ and and $\hat{J}$ are themselves randomly sampled from a probability distribution.

4.1 Top Two Probability Sampling (TTPS)

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

4.2 Top Two Value Sampling (TTVS)

We now propose a variant of top-two sampling which considers not only the probability a design is optimal, but the expected amount by which its value 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 β and the best alternative

$\hat{J}_n$ otherwise. As observations are gathered, beliefs are updated and so the top two designs change over time. The measure of value $V_{n,i}$ is defined below.

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

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

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

$$V_{n,i} = \int_{\Theta} v_i(\boldsymbol{\theta}) \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\Theta_i} v_i(\boldsymbol{\theta}) \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{3}$$

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 value of having the option to choose design i when it is revealed to be the best design. Note that the integral (3) defining $V_{n,i}$ is a weighted version of the integral defining $\alpha_{n,i}$. The paper will formalize a sense in which $V_{n,i}$ and $\alpha_{n,i}$ are asymptotically equivalent as $n \to \infty$, and as a result the asymptotic analysis of Top-Two Value sampling essentially reduces to the analysis of Top Two Probability sampling.

4.3 Thompson Sampling

The Thompson sampling 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 under a parameter drawn from the posterior distribution. Efficient implementations of Thompson sampling typically generate a sample from $(\alpha_{n,1}, ..., \alpha_{n,k})$ by first drawing a sample $\hat{\theta}$ from Π_n and then selecting the action $\arg\max_i \hat{\theta}_i$.

Thompson sampling can have very poor asymptotic performance for the best arm identification problem. Intuitively, the reason for this is that 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 $\alpha_{n,i} > .99$. In fact, 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's 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.

4.4 Pure Exploration Thompson Sampling (PTS)

This section proposes pure-exploration Thompson sampling (PTS), 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 $I \sim \boldsymbol{\alpha}_n$. Design I is measured with probability β, but, in order to prevent the algorithm from exclusively focusing on

one action, with probability $1-\beta$, a different design is measured. This alternative design is drawn from the probability distribution $\boldsymbol{\alpha}_n^{-I}$, which is derived from $\boldsymbol{\alpha}_n$ by setting $\alpha_{n,I}^{-I} = 0$ and re-scaling other entries

$$\alpha_{n,j}^{-I} = \frac{\alpha_{n,j}}{1-\alpha_{n,I}} \qquad j \neq I$$

so that $\boldsymbol{\alpha}_n^{-I}$ forms a valid pmf. Here, $\alpha_{n,j}^{-I}$ captures the probability the best design is j given it is not I. Samples can be drawn from $\boldsymbol{\alpha}_n^{-I}$ by sampling from $\boldsymbol{\alpha}_n$ and using rejection sampling.

Much like standard Thompson sampling, this policy can be implemented without accurately computing the distribution $\boldsymbol{\alpha}_n$. Algorithm 1 shows how to sample an action according to the above procedure by drawing samples of parameters from Π_n. Either through the choice of conjugate prior distributions, or through the use of Markov chain Monte Carlo, it's possible to sample parameters from the posterior distribution for many interesting models.

Algorithm 1 Pure Exploration Thompson Sampling (β)

Sample $\hat{\boldsymbol{\theta}} \sim \Pi_n$ and set $I \leftarrow \arg\max_i \hat{\theta}_i$ ▷ Apply Thompson sampling
Sample $U \sim \text{Uniform}(0,1)$
if $U < \beta$ **then** ▷ Occurs with probability β.
    Play I
else
    repeat
        Sample $\hat{\boldsymbol{\theta}} \sim \Pi_n$ and set $J \leftarrow \arg\max_j \hat{\theta}_j$ ▷ Repeat Thompson sampling
    until $J \neq I$
    Play J
end if

5 A Numerical Experiment

Some of the paper's main insights are reflected in a simple numerical experiment. Consider a problem where observations are binary $Y_{n,i} \in \{0,1\}$, and the unknown vector $\boldsymbol{\theta}^* = (.1, .2, .3, .4, .5)$ defines the true success probability of each design. Each algorithm begins with an independent uniform prior over the components of $\boldsymbol{\theta}^*$. The experiment compares the performance of Top Two Probability sampling (TTPS), Top Two Value sampling (TTVS)[1], and pure exploration Thompson sampling (PTS) with $\beta = 1/2$ against a uniform allocation rule which allocates equal measurement effort ($\psi_{n,i} = 1/5$) to each design. The uniform allocation is a natural benchmark, as it's the most commonly used strategy in practice.

All results are averaged across 150 trials. Figure 1 displays the measurements required for the posterior to reach a given confidence level. In particular, the experiment tracks the first time when $\max_i \alpha_{n,i} \geq .95$, $\max_i \alpha_{n,i} \geq .99$ and $\max_i \alpha_{n,i} \geq .999$. Figure 1 displays the average number of measurements required for each algorithm to reach each fixed confidence level. 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.

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 150 trials. Panel (a) displays

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

the average number of measurements collected from each design. It's striking that although PTS, TTPS, and TTVs seem quite different, they all settle on essentially the same distribution of measurement effort. Because $\beta = 1/2$, roughly one half of the measurements are collected from $I^* = 5$. Moreover, fewer measurements are collected from designs that are farther from optimal, and most of the remaining half of measurement effort is allocated to design 4. Notice that using the same number of noisy samples it's much more difficult certify that $\theta_4^* < \theta_5^*$ than that $\theta_1^* < \theta_5^*$, both because θ_4^* is closer to θ_5^*, and because observations from a Bernoulli distribution with parameter .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, the PTS, TTPS, and TTVS appear to exactly adjust measurement effort to gather *equal evidence* that each of the first four designs is not optimal.

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

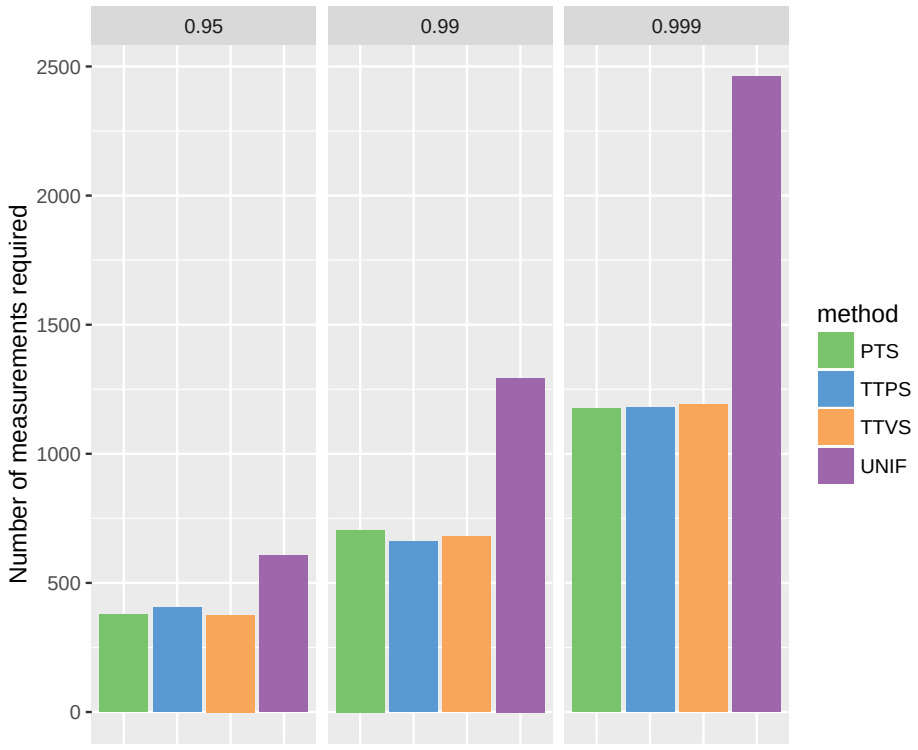
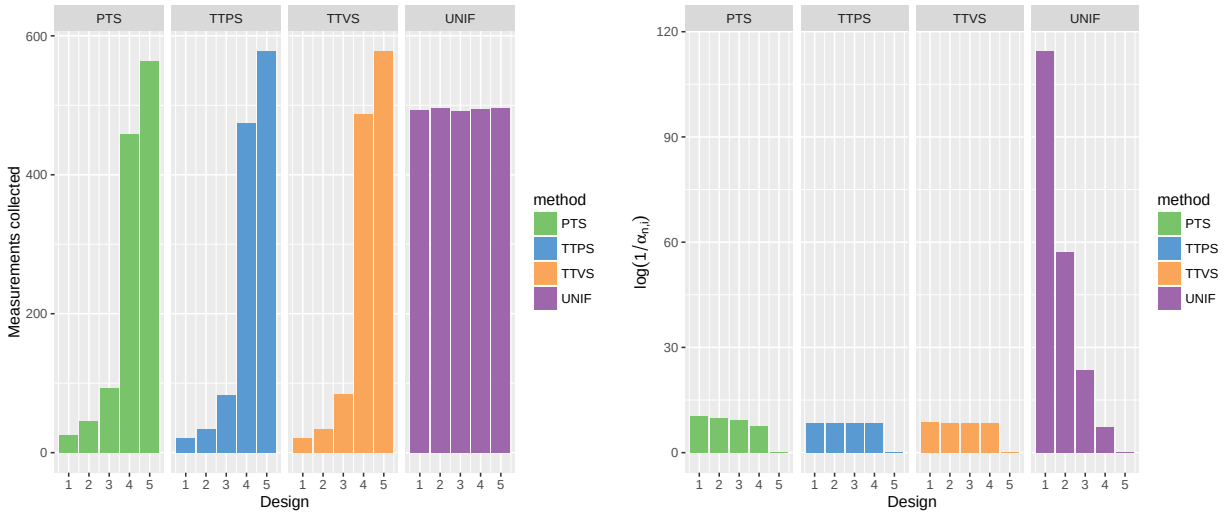


Figure 1: Number of measurements required to reach given confidence level.



(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.

6 Main Theoretical Results

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

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

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

To facilitate theoretical analysis, I will make three additional boundedness assumptions, which are assumed throughout all subsequent theoretical analysis.

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

$$0 < \inf_{\boldsymbol{\theta} \in \Theta} \pi_0(\boldsymbol{\theta}) < \sup_{\boldsymbol{\theta} \in \Theta} \pi_0(\boldsymbol{\theta}) < \infty,$$

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

The next theorem establishes that there is an exponent $\Gamma^* > 0$ such that $\Pi_n(\Theta_{-I^*}) \to 0$ at rate faster than $e^{-n\Gamma^*/2}$ under each proposed algorithm with parameter $\beta = 1/2$, but there is no allocation rule under which convergence is faster than $e^{-n\Gamma^*}$. The exponent Γ^* will be explicitly characterized in Section 7

Theorem 1. *There is a constant* $\Gamma^* > 0$ *such that under any adaptive allocation rule*

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

If PTS, TTPS, or TTVS is executed with parameter $\beta = 1/2$, then

$$\liminf_{n\longrightarrow\infty} \; -\frac{1}{n}\log \Pi_n(\Theta_{-I^*}) \geq \Gamma^*/2.$$

This theorem shows that when $\beta = 1/2$, the exponent established for the proposed algorithms is always within a factor of 2 of the optimal exponent. The paper will in fact exactly characterize the exponent Γ^*_β attained PTS, TTPS, and TTVS with parameter β. In the long run, these algorithms each allocate a faction β of overall effort to measuring I^*, and so in general $\Gamma^*_\beta \neq \Gamma^*$. As shown by the next theorem, however, $\max_\beta \Gamma^*_\beta = \Gamma^*$, so if β is properly tuned, these algorithms will attain the optimal exponent Γ^*. Moreover, they are shown to attain the optimal rate of convergence among algorithms which allocate a faction β of overall effort to measuring I^*. In this sense, they optimally adjust the measurement effort allocated to each of the $k-1$ suboptimal designs.

Theorem 2. *There exist constants $\{\Gamma^*_\beta > 0 : \beta \in (0,1)\}$, satisfying $\Gamma^* = \sup_\beta \Gamma^*_\beta \leq 2\Gamma^*_{\frac{1}{2}}$ and, such that if PTS, TTPS, or TTVS are executed with parameter β, $\overline{\boldsymbol{\psi}}_{n,I^*} \longrightarrow \beta$ and*

$$\lim_{n\longrightarrow\infty} \; -\frac{1}{n}\log \Pi_n(\Theta_{-I^*}) = \Gamma^*_\beta.$$

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

$$\limsup_{n\longrightarrow\infty} \; -\frac{1}{n}\log \Pi_n(\Theta_{-I^*}) \leq \Gamma^*_\beta \qquad a.s.$$

6.1 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 of 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 σ–subgaussian 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 is dominated by that of a zero mean Gaussian random variable with variance σ^2. Gaussian random variables are sub-Gaussian, as are uniformly bounded random variables. The next result applies to both Bernoulli and Gaussian distributions, as each can be parameterized with sufficient statistic $T(y) = y$.

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

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

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

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

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

This shows that $\Pi_n(\Theta_{-I^*})$ decays asymptotically 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. Proposition 1 replaces the dependence on the smallest gap Δ_i with a dependence on $\left(\sum_{i=2}^k \Delta_i^{-2}\right)^{-1}$, which captures

the average inverse gap. This rate is attained only by an intelligent adapative 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 $Y_{n,I^*} \sim N(0, \sigma^2)$ and $Y_{n,i} \sim N(-\Delta_i, \sigma^2)$ for each $i \neq I^*$,*

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

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

7 Analysis

7.1 Asymptotic Notation.

To simplify the presentation, it's useful to develop further 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 $\frac{1}{n} \log(\frac{a_n}{b_n}) \to 0$ as $n \to \infty$. This notation means that a_n and b_n are *equal up to first order in the exponent*. In particular

$$a_n \doteq e^{-cn} \iff a_n = e^{-(c+o(1))n}.$$

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. Together these relations show that if $a_n \in \mathbb{R}^k$ for each n, and $c \in \mathbb{R}^k_{++}$, then

$$\sum_{i=1}^{k} c_i a_{n,i} \doteq \max_{i \in \{1,\ldots,k\}} a_{n,i} \qquad \text{as } n \to \infty$$

When applied to sequences of random variables, these relations are understood to apply almost surely.

7.2 Posterior Large Deviations

This section provides an asymptotic characterization of posterior probabilities of the form

$$\Pi_n(\tilde{\Theta}) = \int_{\tilde{\Theta}} \pi_n(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{4}$$

for any open set $\tilde{\Theta} \subset \Theta$ and under any adaptive measurement strategy.

The solution depends on the notion of Kullback-Leibler divergence. For two parameters $\theta, \theta' \in \mathbb{R}$, the log-likelihood ratio, $\log\left(p(y|\theta)/p(y|\theta')\right)$, provides a measure of the amount of information y provides in favor of θ over θ'. The Kullback-Leibler divergence

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

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

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

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, $\psi_{n,i} \triangleq \mathbf{P}(I_n = i|\mathcal{F}_n)$, changes over time as data is collected. Recall that

$$\overline{\psi}_{n,i} \triangleq n^{-1} \sum_{\ell=1}^{n} \psi_{n,i}$$

captures the fraction of overall effort allocated to measuring design i over the first n periods. Under an adaptive allocation rule, $\overline{\boldsymbol{\psi}}_n$ is function of the history $(I_1, Y_{1,I_1}, ...I_{n-1}, Y_{n-1,I_{n-1}})$ and is therefore a random variable. Given that measurement effort has been allocation according to $\overline{\boldsymbol{\psi}}_n$, $D_{\overline{\boldsymbol{\psi}}_n}(\boldsymbol{\theta}^*||\boldsymbol{\theta})$ quantifies the average information acquired that distinguishes $\boldsymbol{\theta}$ from the true parameter $\boldsymbol{\theta}^*$. The following theorem relates the posterior mass assigned to $\tilde{\Theta}$ to $\inf_{\boldsymbol{\theta}\in\tilde{\Theta}} D_{\overline{\boldsymbol{\psi}}_n}(\boldsymbol{\theta}^*||\boldsymbol{\theta})$, which captures the element in $\tilde{\Theta}$ that is hardest to distinguish from $\boldsymbol{\theta}^*$ based on samples from $\overline{\boldsymbol{\psi}}_n$.

Proposition 3. *For any open set* $\tilde{\Theta} \subset \Theta$*,*

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

To understand this result, consider a simpler problem 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 posteriors densities

$$\log\left(\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)}\right) = \log\left(\frac{\pi_0(\boldsymbol{\theta})}{\pi_0(\boldsymbol{\theta}^*)}\right) + \sum_{\ell=1}^{n} \log\left(\frac{p(Y_{\ell,i}|\theta_i)}{p(Y_{\ell,i}|\theta_i^*)}\right)$$

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 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 $n \longrightarrow \infty$, $n^{-1}\log\left(\pi_n(\boldsymbol{\theta})/\pi_n(\boldsymbol{\theta}^*)\right) \longrightarrow -d(\theta_i^*||\theta_i)$, 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 $(I_1, I_2, I_3, ...)$ are drawn independently of the observations $(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3, ...)$ from a fixed probability distribution $\boldsymbol{\psi}$, then

$$\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} \doteq \exp\left\{-nD_{\boldsymbol{\psi}}(\boldsymbol{\theta}^*||\boldsymbol{\theta})\right\}. \tag{5}$$

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 3, but there are several additional technical challenges involved in a rigorous proof. First, we need that a property like (5) to hold 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 (5). Finally, since $\overline{\boldsymbol{\psi}}_n$ changes over time, the point $\arg\min_{\boldsymbol{\theta}\in\tilde{\Theta}} D_{\overline{\boldsymbol{\psi}}_n}(\boldsymbol{\theta}^*||\boldsymbol{\theta})$ changes over time and basic Laplace approximations don't directly apply.

7.3 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 $\boldsymbol{\theta}^*$, she does not know which measurements will provide the most information. To characterize the optimal exponent Γ^*, however, it's useful to consider the easier problem of gathering the most effective evidence when $\boldsymbol{\theta}^*$ is known. We can cast this as a game between two players:

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

In order to drive the posterior probability α_{n,I^*} to 1, the decision-maker must be able to rule out all parameters in Θ_{-I^*} under which the optimal action is not I^*. Our analysis shows that the posterior probability assigned to Θ_{-I^*} is dominated by the parameter that is hardest to distinguish from $\boldsymbol{\theta}^*$ under $\overline{\boldsymbol{\psi}}_n$. In particular, by Proposition 3,

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

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

$$\max_{\boldsymbol{\psi}} \min_{\boldsymbol{\theta}\in\Theta_{-I^*}} D_{\boldsymbol{\psi}}(\boldsymbol{\theta}^*||\boldsymbol{\theta}) \tag{6}$$

represents an asymptotically optimal allocation rule. As highlighted in the literature review, the max-min problem (6) 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 (6). Much of our analysis involves the posterior probability 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

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

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

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

under which the value at i exceeds that at I^*. Since $\Theta_{-I^*} = \cup_{i \neq I^*} \overline{\Theta}_i$,

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

and therefore

$$\Pi_n(\Theta_{-I^*}) \doteq \max_{i \neq I^*} \Pi_n(\overline{\Theta}_i) \tag{7}$$

This yields an analogue of (6) that will simplify our subsequent analysis. Combining (7) with Proposition 3 shows the solution to the max-min problem

$$\Gamma^* \triangleq \max_{\boldsymbol{\psi}} \min_{i \neq I^*} \min_{\boldsymbol{\theta} \in \overline{\Theta}_i} D_{\psi}(\boldsymbol{\theta}^* || \boldsymbol{\theta}) \tag{8}$$

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

$$C_i(\beta, \psi) \triangleq \min_{x \in \mathbb{R}} \beta d(\theta^*_{I^*} || x) + \psi d(\theta^*_{i^*} || x). \tag{9}$$

The next lemma shows that the function C_i arises as the solution to the minimization problem over $\boldsymbol{\theta} \in \overline{\Theta}_i$ in (8). It also shows that the minimum in (9) is attained by a parameter $\overline{\theta}$ under which the mean observation is a weighted combination of the means of $\theta^*_{I^*}$ and θ^*_i. Recall that, for an exponential family distribution $A'(\theta) = \int T(y) p(y|\theta) d\nu(y)$ is the mean observation of the sufficient statistic $T(y)$ under θ.

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

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

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

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

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

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

Lemma 1 immediately implies

$$\Gamma^* = \max_{\boldsymbol{\psi}} \min_{i \neq I^*} C_i(\psi_{I^*}, \psi_i). \tag{10}$$

The function $C_i(\beta, \psi)$ captures the effectiveness with which one can certify $\theta^*_{I^*} \geq \theta^*_i$ using an allocation rule which measures actions I^* and i with respective frequencies β and ψ. The function $C_i(\beta, \psi)$ is an increasing function of the measurement effort (β, ψ) allocated to designs I^* and i. For given β and ψ, $C_i(\beta, \psi) \geq C_j(\beta, \psi)$ when $\theta^*_i \leq \theta^*_j$, reflecting that θ^*_i is easier to distinguish from $\theta^*_{I^*}$ than θ^*_j.

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 $\psi_{n,i} \triangleq \mathbf{P}(I_n = i | \mathcal{F}_n)$ denotes the measurement effort assigned design i at time n.

Proposition 4. *Let $\boldsymbol{\psi}*$ denote the optimal solution to the maximization problem (10). If $\boldsymbol{\psi}_n = \boldsymbol{\psi}^*$ for all n, then*

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

Moreover under any other adaptive allocation rule,

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

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

An Optimal Constrained Allocation. Because the algorithms studied in this paper always allocate β–fraction of their samples to measuring I^* in the long run, they may not exactly attain the optimal error exponent. To make rigorous claims about their performance, consider a modified version of the error exponent (10) given by the constrained max-min problem

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

This optimization problem yields the optimal allocation subject to a constraint that β–fraction of the samples are spent on I^*. The next subsection will show that PTS, TTPS, and TTVS attain the error exponent Γ^*_β. The next proposition formalizes that the solution to this optimization problem represent 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 $C_2(\beta, \psi_2) < C_j(\beta, \psi_3)$, 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's optimal to decrease ψ_2 and increase ψ_1, until the point when $C_2(\beta, \psi_2) = C_3(\beta, \psi_3)$. Instead of allocating *equal measurement effort* to each alternative, its 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 5. *The solution to the optimization problem (11) is the unique allocation ψ^* satisfying $\psi^*_{I^*} = \beta$ and*

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

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

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

Moreover under any other adaptive allocation rule, if $\psi_{n,I^*} = \beta$ for all n then

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

almost surely.

The following Lemma relates the constrained explonent Γ^*_β to Γ^*.

Lemma 2. $\max_{\beta\in[0,1]} \Gamma^*_\beta = \Gamma^*$ *and* $\Gamma^*_{\frac{1}{2}} \geq \Gamma^*/2$.

7.4 Optimal Adaptive Allocation: the TTPS, TTVS, and PTS Algorithms

While the last subsection describes an asymptotically optimal exploration strategy, implementing this strategy requires knowledge of the parameter vector $\boldsymbol{\theta}^*$. One natural approach to attaining the rate (10) would be to alternate between collecting data to construct a point estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$ and solving the optimization problem (10) with $\hat{\boldsymbol{\theta}}$ in place of of $\boldsymbol{\theta}^*$. While such an approach likely works well asymptotically, it may perform very poorly in early stages when $\hat{\boldsymbol{\theta}}$ is a poor estimate of $\boldsymbol{\theta}^*$.

Instead of attempting to directly solve the optimization problem (10), this paper focuses on simple and intuitive heuristic strategies, which, ostensibly, have no connection to the derivations shown above. By investigating the structure of the optimal exponent, and the dynamics of pure-exploration Thompson sampling, Top-Two Probability sampling and Top-Two Value Sampling, we show that all three methods automatically compete with the unknown optimal allocation.

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

Proposition 6. *Under the PTS, TTPS, or TTVS algorithm with parameter $\beta > 0$,*

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

The proof proceeds by appealing to Proposition 5 and showing $\overline{\boldsymbol{\psi}}_n \longrightarrow \boldsymbol{\psi}^\beta$, where $\boldsymbol{\psi}^\beta$ is the unique allocation with $\psi^\beta_{I^*} = \beta$ satisfying

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

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. The proof of this lemma is given in Section B.3.

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

This 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 $n \to \infty$, since it tends to zero at an exponential rate in n whereas $v_i(\boldsymbol{\theta})$ is a fixed function of n.

To understand Proposition 6, imagine that n is very large, and the algorithm has allocated too much measurement effort to a suboptimal action i, so $\overline{\psi}_{n,i} > \psi^\beta_i + \delta$ for a constant $\delta > 0$. Then the algorithm has allocated too little measurement effort to at least one other suboptimal design $j \neq i$. Then, since much less evidence has been gathered about j than i, we expect $\alpha_{j,n} >> \alpha_{j,i}$. When this occurs, PTS, TTPS and TTVS essentially never sample action i until the average effort $\overline{\psi}_{n,i}$ allocated to design i dips back down toward ψ^β_i. This seems to suggest that the algorithm cannot allocate too much effort to any alternative, but that in turn implies that it never allocates too little effort to measuring any alternative.

Acknowledgements: I would like to thank Sebastien Bubeck and Peter Glynn for helpful discussions related to this work.

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.

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.

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.

Chun-Hung Chen, Jianwu Lin, Enver Yücesan, and Stephen 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.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory*, pages 255–270. Springer, 2002.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7:1079–1105, 2006.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

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.

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.

Kevin Jamieson and Robert 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.

Kris Johnson, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Available at SSRN*, 2015.

Zohar Karnin, Tomer Koren, and Oren 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, 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.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251, 2013.

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.

Shie Mannor and John N Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 2014a. URL http://dx.doi.org/10.1287/moor.2014.0650.

Dan Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pages 1583–1591, 2014b.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. *arXiv preprint arXiv:1511.05219*, 2015.

S.L. Scott. Overview of content experiments: Multi-armed bandit experiments, 2014. URL https://support.google.com/analytics/answer/2844870?hl=en. [Online; accessed 9-February-2014].

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.

David Williams. *Probability with martingales*. Cambridge university press, 1991.

A 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 $X_n \to \infty$ whenever $Y_n \to \infty$, this means that the set $\{\omega : Y_n(\omega) \to \infty, X_n(\omega) \nrightarrow \infty\}$ has measure zero.

Some Martingales Convergence Facts. We first state a law of large numbers for Martingale sequences.

Fact 1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a real valued martingale difference sequence. If*

$$n^{-2}\sum_{\ell=1}^{n}\mathbf{E}\left[X_\ell^2\right] \longrightarrow 0,$$

then

$$n^{-1}\sum_{\ell=1}^{n}X_\ell \longrightarrow 0 \qquad a.s.$$

A corollary of this lemma relates the long run measurement effort $\sum_{\ell=1}^{n}\psi_{n,i}$ to the number of times alternative i is actually measured $S_{n,i} = \sum_{\ell=1}^{n}\mathbf{1}(I_n = i)$.

Corollary 1. *For each alternative $i \in \{1, ..., k\}$,*

$$\frac{1}{n}\sum_{\ell=1}^{n}(\psi_{\ell,i} - \mathbf{1}(I_\ell = i)) \longrightarrow 0$$

The following result shows that in addition $S_n \to \infty$ if and only if $\sum_{\ell=1}^n \psi_{n,i} \to \infty$. This is Levy's conditional form of the Borel-Cantelli lemmas.

Fact 2. *(Williams [1991], 12.15) Let $(X_n : n \in \mathbb{N})$ be a sequence of binary random variables adapted to the filtration $(\mathcal{H}_n : n = 0, 1, 2...)$. Then*

$$\lim_{n\to\infty} \sum_{\ell=1}^{n} X_\ell = \infty \iff \lim_{n\to\infty} \sum_{\ell=1}^{n} \mathbf{P}(X_\ell = 1|\mathcal{H}_{\ell-1}) = \infty.$$

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

$$A'(\theta) = \int y p(y|\theta) d\nu(y) \tag{12}$$

equal to the mean under θ. The Kullback-Leibler divergence is equal to

$$d(\theta||\theta') = (\theta - \theta')A'(\theta) - A(\theta) + A(\theta') \tag{13}$$

and satisfies

$$\theta'' > \theta' \geq \theta \implies d(\theta||\theta'') > d(\theta||\theta') \tag{14}$$
$$\theta'' < \theta' \leq \theta \implies d(\theta||\theta'') < d(\theta||\theta') \tag{15}$$

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

$$\sup_{\theta\in[\underline{\theta},\overline{\theta}]} |A(\theta)| < \infty \qquad \& \qquad \sup_{\theta,\theta'\in[\underline{\theta},\overline{\theta}]} d(\theta||\theta') < \infty. \tag{16}$$

This effectively rules out cases where a single observation can provide enough information to completely rule out a parameter.

B Concentration of the Posterior Distribution

B.1 Large Deviations: Proof of Proposition 3

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

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

Note that $W_n(\boldsymbol{\theta}^*) = 0$. Define

$$\Lambda_n(\boldsymbol{\theta}) \triangleq \log\left(\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)}\right)$$

and

$$\lambda_0(\boldsymbol{\theta}) \triangleq \log\left(\frac{\pi_0(\boldsymbol{\theta})}{\pi_0(\boldsymbol{\theta}^*)}\right) \qquad \lambda_n(\boldsymbol{\theta}) \triangleq \log\left(\frac{p(Y_{n,I_n}|\theta_{I_n})}{p(Y_{n,I_n}|\theta^*_{I_n})}\right) \quad n \in \mathbb{N}$$

so that

$$\Lambda_n(\boldsymbol{\theta}) = \sum_{\ell=0}^{n} \lambda_\ell(\boldsymbol{\theta}).$$

One can use Fact 1 to show that $\frac{\pi_n(\boldsymbol{\theta})}{\pi_n(\boldsymbol{\theta}^*)} \doteq \exp\{-nW_n(\boldsymbol{\theta})\}$. The next lemma provides a stronger result, and shows that such a relation holds uniformly over $\boldsymbol{\theta}$.

Lemma 4.

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{\Lambda_n(\boldsymbol{\theta})}{n}+W_n(\boldsymbol{\theta})\right|\longrightarrow 0$$

Proof.

$$\Lambda_n(\boldsymbol{\theta})=\sum_{\ell=0}^{n}\lambda_\ell(\boldsymbol{\theta})$$

For each $n\geq 1$

$$\mathbf{E}[\lambda_n(\boldsymbol{\theta})|\mathcal{F}_n,I_n]=d(\theta^*_{I_n}||\theta_{I_n}).$$

This implies

$$\mathbf{E}[\lambda_n(\boldsymbol{\theta})|\mathcal{F}_n]=\sum_{i=1}^{k}\mathbf{P}(I_n=i|\mathcal{F}_n)d(\theta_i^*||\theta_i)=\sum_{i=1}^{k}\psi_{n,i}d(\theta_i^*||\theta_i)$$

and therefore that,

$$W_n(\boldsymbol{\theta})=\frac{1}{n}\sum_{\ell=1}^{n}\mathbf{E}[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell].$$

Now, because the variance of the increments $\lambda_\ell(\boldsymbol{\theta})-\mathbf{E}[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell]$ is bounded uniformly, the law of large numbers for martingale difference sequences shows

$$\frac{1}{n}\sum_{\ell=1}^{n}\left(\lambda_\ell(\boldsymbol{\theta})-\mathbf{E}[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell]\right)\longrightarrow 0.$$

Since $\lambda_0(\boldsymbol{\theta})/n\to 0$, this would imply that $\frac{1}{n}\Lambda_n(\boldsymbol{\theta})-W_n(\boldsymbol{\theta})\longrightarrow 0$ for every $\boldsymbol{\theta}$. To show uniform convergence, we leverage the additional structure provided by our assumptions about the prior and likelihood. Write,

$$\frac{1}{n}\Lambda_n(\boldsymbol{\theta})-W_n(\boldsymbol{\theta})=\underbrace{\frac{\lambda_0(\boldsymbol{\theta}))}{n}}_{(a)}+\underbrace{\frac{\sum_{\ell=1}^{n}\left(E[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell,I_\ell]-\mathbf{E}[\lambda_n(\boldsymbol{\theta})|\mathcal{F}_\ell]\right))}{n}}_{(b)}+\underbrace{\frac{\sum_{\ell=1}^{n}\left(\lambda_\ell(\boldsymbol{\theta})-\mathbf{E}[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell,I_\ell]\right))}{n}}_{(c)}$$

We show that each of these terms converges to zero uniformly in $\boldsymbol{\theta}$.
(a): Since

$$\lambda_0(\boldsymbol{\theta})\triangleq\log\left(\frac{\pi_0(\boldsymbol{\theta})}{\pi_0(\boldsymbol{\theta}^*)}\right)$$

and we have assumed $\pi_0(\boldsymbol{\theta})$ is uniformly bounded, $n^{-1}\lambda_0(\boldsymbol{\theta})\to 0$ uniformly in $\boldsymbol{\theta}$.
(b): For each $n\in\mathbb{N}$,

$$E[\lambda_n(\boldsymbol{\theta})|\mathcal{F}_n,I_n]-\mathbf{E}[\lambda_n(\boldsymbol{\theta})|\mathcal{F}_n])=\sum_{i=1}^{k}(\mathbf{1}(I_n=i)-\psi_{n,i})d(\theta_i^*||\theta_i)$$

where, as shown in Section A, $d(\theta_i^*||\theta_i)$ is bounded uniformly by

$$C=\max_{1\leq i\leq k}\max_{\theta'\in[\underline{\theta},\bar{\theta}]}d(\theta_i^*||\theta_i')<\infty.$$

We therefore have by Corollary 1 that for each $i=1,..,k$,

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{\ell=1}^{n}\left(\mathbf{E}[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell,I_\ell]-\mathbf{E}[\lambda_n(\boldsymbol{\theta})|\mathcal{F}_\ell]\right)\right|\leq C\sum_{i=1}^{k}\left|\frac{1}{n}\sum_{\ell=1}^{n}(\mathbf{1}(I_\ell=i)-\psi_{n,i})\right|\longrightarrow 0.$$

(c)

$$\lambda_\ell(\boldsymbol{\theta}) = \sum_{i=1}^{k} \mathbf{1}(I_\ell = i)\left((\theta_i - \theta_i^*)T(Y_{\ell,i}) - (A(\theta_i) - A(\theta_i^*)\right)$$

so

$$\begin{aligned}\lambda_\ell(\boldsymbol{\theta}) - \mathbf{E}[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell, I_\ell] &= \sum_{i=1}^{k} \mathbf{1}(I_\ell = i)\left((\theta_i - \theta_i^*)(T(Y_{\ell,i}) - \mathbf{E}[T(Y_{\ell,i})|\mathcal{F}_\ell, I_\ell]\right) \\ &= \sum_{i=1}^{k} \mathbf{1}(I_\ell = i)\left((\theta_i - \theta_i^*)(T(Y_{\ell,i}) - \mathbf{E}[T(Y_{\ell,i})]\right)\end{aligned}$$

where the equality follows from the fact that $Y_{\ell,i}$ is independent of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ...\boldsymbol{Y}_{\ell-1}$ and hence independent of the history of observations $\mathcal{F}_\ell$ and the chosen action I_ℓ. Note that

$$|\theta_i - \theta_i^*| \leq \overline{\theta} - \underline{\theta} \triangleq C < \infty$$

is bounded independently of θ_i. This yields

$$\begin{aligned}\left|\frac{\sum_{\ell=1}^{n}\left(\lambda_\ell(\boldsymbol{\theta}) - \mathbf{E}[\lambda_\ell(\boldsymbol{\theta})|\mathcal{F}_\ell, I_\ell]\right)\right)}{n}\right| &= \sum_{i=1}^{k} \frac{\sum_{\ell=1}^{n} \mathbf{1}(I_\ell = i)\left((\theta_i - \theta_i^*)(T(Y_{\ell,i}) - \mathbf{E}[T(Y_{\ell,i})]\right)}{n} \\ &\leq C\sum_{i=1}^{k}\left|\frac{\sum_{\ell=1}^{n} \mathbf{1}(I_\ell = i)\left(T(Y_{\ell,i}) - \mathbf{E}[T(Y_{\ell,i})]\right)}{n}\right|.\end{aligned}$$

It's then enough that for each $i \in \{1, ..., k\}$,

$$\frac{1}{n}\sum_{\ell=1}^{n} \mathbf{1}(I_\ell = i)\left(T(Y_{\ell,i}) - \mathbf{E}[Y_{\ell,i}]\right) \longrightarrow 0. \tag{17}$$

Because this is a martingale difference sequence with

$$\sup_{n\in\mathbb{N}} \mathbf{E}[(\mathbf{1}(I_n = i)(T(Y_{n,i}) - \mathbf{E}[Y_{n,i}])^2] \leq \sup_{n\in\mathbb{N}} \mathbf{E}[(T(Y_{n,i}) - \mathbf{E}[Y_{n,i}])^2] = \mathrm{Var}(T(Y_{1,i})) < \infty,$$

the limit (17) follows from Fact 1. □

Lemma 5. *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 the previous lemma, we can fix a sequence $\epsilon_n \geq 0$ with $\epsilon_n \to 0$ 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

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

□

Lemma 6. *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

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

This is always possible, 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

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

We have

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

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

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

We now show the reverse. Fix an arbitrary $\epsilon > 0$. It follows from the uniform bound on $A'(\theta)$ (See Lemma 7 below), that there exists $\delta > 0$ such that

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

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

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

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

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

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

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

Taking the logarithm of both sides implies

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

Since ϵ was chosen arbitrarily, this shows

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

and completes the proof. □

Lemma 7. *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

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

where $C = \sup_\theta |A'(\theta)| < \infty$. □

B.2 Posterior Consistency

Proposition 7. *With probability 1, if $S_{n,i} \to \infty$ then for all $\epsilon > 0$*

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

as $n \to \infty$. If $S_{n,i} \nrightarrow \infty$, then

$$\inf_{n\in\mathbb{N}} \Pi_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta', \theta'')\}) > 0 \qquad \forall\ (\theta', \theta'') \subset (\underline{\theta}, \overline{\theta})$$

This proposition says that if design i is measured infinitely often, the marginal posterior distribution of its quality concentrates around the true value θ_i^*. The second part is an anti-concentration result. It says that when design i is not measured infinitely often, the marginal posterior of its quality doesn't concentrate around any value.

Because we don't assume an independent prior across the designs, Π_0 is not a product measure and therefore neither is Π_n. However, since the prior density is bounded, Π_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

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

with the convention that $L_{n,i}(\theta_i) = 1$ when $S_{n,i} = 0$. Therefore $L_n(\boldsymbol{\theta})$ forms the density of a product measure. By normalizing, this induces a probability measure over Θ, which, as we argue in the next lemma, behaves like the posterior Π_n.

Lemma 8. *Set* $C = \sup_{\theta \in [\underline{\theta}, \overline{\theta}]} \pi_0(\theta) < \infty$. *Then for any set* $\tilde{\Theta} \subset \Theta$

$$C^{-1} \frac{\int_{\tilde{\Theta}} L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}} \leq \Pi_n(\tilde{\Theta}) \leq C \frac{\int_{\tilde{\Theta}} L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

In addition, for any interval $[\theta', \theta''] \subset [\underline{\theta}, \overline{\theta}]$ *and* $i \in \{1, .., k\}$,

$$C^{-1} \frac{\int_{\theta'}^{\theta''} L_{n,i}(\theta) d\theta}{\int_{\underline{\theta}}^{\overline{\theta}} L_{n,i}(\theta) d\theta} \leq \Pi_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta', \theta'')\}) \leq C \frac{\int_{\theta'}^{\theta''} L_{n,i}(\theta) d\theta}{\int_{\underline{\theta}}^{\overline{\theta}} L_{n,i}(\theta) d\theta}.$$

Proof. The first step follows immediately since

$$\Pi_n(\tilde{\Theta}) = \frac{\int_{\tilde{\Theta}} \pi_0(\boldsymbol{\theta}) L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\Theta} \pi_0(\boldsymbol{\theta}) L_n(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

The next step follows because L_n is the density of a product measure, and $\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta', \theta'')\}$ is the cartesian product of intervals. □

We can now prove Proposition 7.

Proof of Proposition 7. We begin with the first part of the result. For simplicity of notation, we focus on the upper interval $\tilde{\Theta} = \{\boldsymbol{\theta} \in \Theta : \theta_i > \theta^* + \epsilon\}$, but results follow identically for the lower interval. We want to show $\Pi_n(\tilde{\Theta}) \to 0$, which occurs if and only if $\log \Pi_n(\tilde{\Theta}) \to -\infty$. We have

$$\Pi_n(\tilde{\Theta}) \leq C \frac{\int_{\theta_i^* + \epsilon}^{\overline{\theta}} L_{n,i}(\theta) d\theta}{\int_{\underline{\theta}}^{\overline{\theta}} L_{n,i}(\theta) d\theta} \leq C \frac{\int_{\theta_i^* + \epsilon}^{\overline{\theta}} L_{n,i}(\theta) d\theta}{\int_{\theta_i^*}^{\theta_i^* + \epsilon/2} L_{n,i}(\theta) d\theta} \leq C' \sup_{\substack{\theta > \theta^* + \epsilon \\ \theta' \in [\theta^*, \theta^* + \epsilon/2]}} \frac{L_{n,i}(\theta)}{L_{n,i}(\theta')}$$

where $C' = C(\overline{\theta} - \theta_i^* - \epsilon)/(\epsilon/2)$ is independent of n. It therefore suffices to show that $L_{n,i}(\theta)/L_{n,i}(\theta') \to 0$ uniformly over $\theta > \theta_i^*$ and $\theta' \in [\theta_i^*, \theta_i^* + \epsilon/2]$. By Corollary 2 (which is shown below),

$$S_{n,i}^{-1} \log \left(\frac{L_{n,i}(\theta)}{L_{n,i}(\theta')} \right) \longrightarrow d(\theta^* || \theta) - d(\theta^* || \theta')$$

uniformly in θ, θ'. Since

$$\sup_{\substack{\theta > \theta^* + \epsilon \\ \theta' \in [\theta^*, \theta^* + \epsilon/2]}} d(\theta^* || \theta) - d(\theta^* || \theta') = d(\theta^* || \theta^* + \epsilon/2) - d(\theta^* || \theta^* + \epsilon) < 0$$

this implies

$$\log \left(\frac{L_{n,i}(\theta)}{L_{n,i}(\theta')} \right) = S_{n,i} \left(S_{n,i}^{-1} \log \left(\frac{L_{n,i}(\theta)}{L_{n,i}(\theta')} \right) \right) \to -\infty$$

uniformly in θ, θ' in this range, which implies the result.

The second part of the claim follows immediately from the lower bound in Lemma 8 of

$$\Pi_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta', \theta'')\}) \geq C^{-1} \frac{\int_{\theta'}^{\theta''} L_{n,i}(\theta) d\theta}{\int_{\underline{\theta}}^{\overline{\theta}} L_{n,i}(\theta) d\theta}. \tag{18}$$

If $\lim_{n \to \infty} S_{n,i} < \infty$ then there is a finite time τ after which action i is never sampled again. Then, for $n \geq \tau$

$$L_{n,i}(\theta) = L_{\tau,i}(\theta) > 0 \ \forall \theta$$

is bounded independently of n, which implies by (18) that $\inf_{n \in \mathbb{N}} \Pi_n(\{\boldsymbol{\theta} \in \Theta | \theta_i \in (\theta', \theta'')\}) > 0$. □

Lemma 9. *If $S_{n,i} \longrightarrow \infty$, then*

$$\frac{1}{S_{n,i}} \log\left(\frac{L_{n,i}(\theta_i^*)}{L_{n,i}(\theta)}\right) \longrightarrow d(\theta_i^*||\theta).$$

uniformly over $\theta \in [\underline{\theta}, \overline{\theta}]$.

Proof. Let $\mu_i = \mathbf{E}[T(Y_{1,i})]$ denote the expected value of $Y_{1,i}$ under θ_1^* and let $\hat{\mu}_{n,i} = (S_{n,i})^{-1} \sum_{\ell=1}^{n} \mathbf{1}(I_\ell = i) T(Y_{\ell,i})$ denote the empirical average among samples from design i. Then, $\hat{\mu}_{n,i} \longrightarrow \mu_i$ almost surely on the event $S_{n,i} \longrightarrow \infty$. Direct calculation using the density of the exponential family shows

$$\begin{aligned} S_{n,i}^{-1} \log\left(\frac{L_{n,i}(\theta_i^*)}{L_{n,i}(\theta)}\right) = S_{n,i}^{-1} \sum_{\substack{\ell \le n \\ I_\ell = i}} \log\left(\frac{p(Y_{\ell,1}|\theta_1^*)}{p(Y_{\ell,1}|\theta)}\right) &= (\theta_i^* - \theta)\hat{\mu}_{n,i} - A(\theta_i^*) + A(\theta)) \\ &= (\theta_i^* - \theta)\mu_i - A(\theta_i^*) + A(\theta)) + (\theta_i^* - \theta)(\hat{\mu}_{n,i} - \mu_i) \\ &= d(\theta_i^*||\theta) + (\theta_i^* - \theta)(\hat{\mu}_{n,i} - \mu_i) \end{aligned}$$

This completes the proof since

$$|(\theta_1^* - \theta)(\hat{\mu}_{n,i} - \mu_i)| \le |\overline{\theta} - \underline{\theta}||\hat{\mu}_{n,i} - \mu_i)| \longrightarrow 0$$

uniformly in θ as $n \longrightarrow \infty$. □

Corollary 2. *If $S_{n,i} \longrightarrow \infty$, then*

$$S_{n,i}^{-1} \log\left(\frac{L_{n,i}(\theta')}{L_{n,i}(\theta)}\right) \longrightarrow d(\theta_1^*||\theta) - d(\theta_1^*||\theta')$$

uniformly over $\theta, \theta' \in [\underline{\theta}, \overline{\theta}]$.

Proof.

$$S_{n,i}^{-1} \log\left(\frac{L_{n,i}(\theta')}{L_{n,i}(\theta)}\right) = S_{n,i}^{-1} \log\left(\frac{L_{n,i}(\theta^*)}{L_{n,i}(\theta)}\right) - S_{n,i}^{-1} \log\left(\frac{L_{n,i}(\theta^*)}{L_{n,i}(\theta')}\right)$$

and so the result follows from the previous lemma. □

B.3 Asymptotics of the Value Measure: Proof of Lemma 3

First, since

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

it is immediate that

$$\limsup \frac{1}{n}(\log V_{n,i} - \log \alpha_{n,i}) \le 0. \tag{19}$$

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

$$\Theta_{i,\delta} = \{\boldsymbol{\theta} \in \Theta : \theta_i \ge \max_{j \ne i} \theta_j + \delta\}$$

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

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

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

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

Combining this with Proposition 3 shows

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

By Lemma 7, for any $\epsilon > 0$, there exists $\delta > 0$ such that for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and $n \in \mathbb{N}$,

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_\infty \leq \delta \implies D_{\overline{\psi}_n}(\boldsymbol{\theta}^* || \boldsymbol{\theta}) \leq \epsilon.$$

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

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

This shows $\liminf n^{-1}(\log V_{n,i} - \log \alpha_{n,i}) \geq -\epsilon$ for all $\epsilon > 0$. Together with (19) this implies $n^{-1}(\log V_{n,i} - \log \alpha_{n,i}) \to 0$.

C Simplifying and Bounding the Error Exponent

C.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 A that $A(\theta)$ is increasing and strictly convex, and, by (12), $A'(\theta)$ is the mean observation under θ.

Lemma 1. *Define for each $c, \psi \geq 0$,*

$$C_i(\beta, \psi) \triangleq \min_{x \in \mathbb{R}} \beta d(\theta^*_{I^*} || x) + \psi d(\theta^*_i || x). \tag{20}$$

(a) *For any $i \in \{1, .., k\}$ and probability distribution $\boldsymbol{\psi}$ over $\{1, ..., k\}$*

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

(b) *Each C_i is a concave function.*
(c) *The unique solution to the minimization problem (20) is $\overline{\theta} \in \mathbb{R}$ satisfying*

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

Therefore

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

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

Proof. **(a)**

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

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 (θ_i, θ_{I^*}). Then, by the monotonicity properties of KL-divergence (see Section A, equation (14)), there is always a minimum with $\theta_i = \theta_{I^*}$. Therefore this objective value is equal to

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

(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 (β, ψ).

(c) Direct calculation using the formula for KL divergence in exponential families (see (13) in Section A) 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) I will show C_i is strictly increasing in the second argument. The proof that it's 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 ψ_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. □

C.2 Proof of Proposition 5

I will begin by restating Proposition 5.

Proposition 5. *The solution to the optimization problem (11) is the unique allocation ψ^* satisfying $\psi_{I^*}^* = \beta$ and*

$$C_i(\beta, \psi_i) = C_j(\beta, \psi_j) \qquad \forall\, i, j \neq I^*. \tag{21}$$

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

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

Moreover under any other adaptive allocation rule, if $\psi_{n,I^*} = \beta$ for all n then

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

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 $\boldsymbol{\psi}^*$ to (11), so

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

Suppose ψ^* does not satisfy (21), 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 $\boldsymbol{\psi}^\epsilon$ with $\boldsymbol{\psi}^\epsilon_j = \psi^*_j - \epsilon$ and $\psi^\epsilon_i = \psi^*_i + \epsilon/(k-2)$ for each $i \notin \{I^*, j\}$. For sufficiently small ϵ, one has

$$C_j(\beta, \psi^\epsilon_j) > \min_{i\neq I^*} C_i(\beta, \psi^\epsilon_i) > \min_{i\neq I^*} C_i(\beta, \psi^*_i)$$

and so $\boldsymbol{\psi}^\epsilon$ attains a higher objective value. To show the solution to (21) must be unique, imagine $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$ both satisfy (21) 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 (21) 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 that $\sum_{j\neq I^*} \psi_j = \sum_{j\neq I^*} \psi'_j = 1 - \beta$.

The remaining claims follow immediately from Propoosition 3 and Lemma 1, which together show that under any adaptive allocation rule

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

□

C.3 Proof of Lemma 2

Proof. We need to show $\max_{\beta\in[0,1]} \Gamma^*_\beta = \Gamma^*$ and $\Gamma^*_{\frac{1}{2}} \geq \Gamma^*/2$. The first claim follows immediately. To show the second claim, define for each non-negative vector $\boldsymbol{\psi}$, $f(\boldsymbol{\psi}) = \min_{i\neq I^*, \boldsymbol{\theta}\in\Theta_i} C_i(\psi_{I^*}, \psi_i)$. Then

$$\Gamma^* = \max\{f(\boldsymbol{\psi}) : \sum_{i=1}^k \psi_i = 1, \boldsymbol{\psi} \geq 0\} \qquad \Gamma^*_{\frac{1}{2}} = \max\{f(\boldsymbol{\psi}) : \psi_{I^*} = \frac{1}{2}, \sum_{i=1}^k \psi_i = 1, \boldsymbol{\psi} \geq 0\}.$$

Let $\boldsymbol{\psi}^*$ denote a feasible solution with $f(\boldsymbol{\psi}^*) = \Gamma^*$.

Because each C_i is non-decreasing, f is non-decreasing. Since the minimum over $\boldsymbol{\theta}$ 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 $c \geq 1$. We have

$$\begin{aligned}
2\Gamma^*_{\frac{1}{2}} &= \max\{f(2\boldsymbol{\psi}) : \psi_{I^*} = \frac{1}{2}, \sum_{i=1}^{k} \psi_i = 1, \boldsymbol{\psi} \geq 0\} \\
&= \max\{f(\boldsymbol{\psi}) : \psi_{I^*} = 1, \sum_{i=1}^{k} \psi_i = 2, \boldsymbol{\psi} \geq 0\} \\
&\geq \max\{f(\boldsymbol{\psi}) : \psi_{I^*} = \psi^*_{I^*}, \sum_{i=1}^{k} \psi_i = 2, \boldsymbol{\psi} \geq 0\} \\
&\geq \max\{f(\boldsymbol{\psi}) : \psi_{I^*} = \psi^*_{I^*}, \sum_{i=1}^{k} \psi_i = 1, \boldsymbol{\psi} \geq 0\} \\
&= \Gamma^*
\end{aligned}$$

where the second inequality uses that f is non-decreasing. □

C.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 3. *Fix two probability measures* $\mathbf{P}$ *and* $\mathbf{Q}$ *defined on a common measureable space* $(\Omega, \mathcal{F})$. *Suppose that* $\mathbf{P}$ *is absolutely continuous with respect to* $\mathbf{Q}$. *Then*

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

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 3, which provides and analogous lower bound on the KL-divergences when distributions are sub-Gaussian. Recall that, $\mu(\theta) = \int y p(y|\theta) d\nu(y)$ denotes the mean observation under θ.

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

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

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 3 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 $\log \mathbf{E}_{\theta'}\left[\exp\{X\}\right] \leq \lambda^2\sigma^2/2$. Therefore, Fact 3 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. □

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,

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

Let $\mu_{I^*} = A'(\theta^*_{I^*})$ and $\mu_i = A'(\theta^*_i)$ denote the means of designs I^*, i so $\Delta_i = \mu_{I^*} - \mu_i$. By Lemma 1,

$$C_i(1/2, \psi_i) = (1/2)d(\theta^*_{I^*}||\overline{\theta}) + \psi_i d(\theta^*_i||\overline{\theta}).$$

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

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

For $\psi_i \leq 1/2$, the mean under $\overline{\theta}$ is at least

$$\frac{\mu_{I^*} + \mu_i}{2} = \mu_i + \Delta_i/2.$$

Now, using Corollary 3 and the non-negativity of KL-divergence

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

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

$$\psi_i = \frac{1}{2}\left(\sum_2^k \Delta_i^{-2}\right)^{-1} \Delta_i^{-2}$$

yields

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

□

C.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 θ^*_i By Proposition 4, we have

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

By Lemma 1,

$$C_i(k^{-1}, k^{-1}) = k^{-1}d(\theta^*_{I^*}||\overline{\theta}) + k^{-1}d(\theta^*_i||\overline{\theta})$$

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

$$C_i(k^{-1}, k^{-1}) = \frac{2(\theta^*_{I^*} - \overline{\theta})^2}{2\sigma^2} = \frac{(\theta^*_{I^*} - \theta^*_i)^2}{4\sigma^2} = \frac{\Delta_i^2}{4\sigma^2}.$$

□

D Analysis of PTS and TTS Allocation Rules: Proof of Proposition 6

Instead of proving Proposition 6 directly, we establish the following result.

Lemma 10. *$\overline{\boldsymbol{\psi}}_n \longrightarrow \boldsymbol{\psi}^\beta$, where $\boldsymbol{\psi}^\beta$ is the unique allocation with $\psi^\beta_{I^*} = \beta$ satisfying*

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

Because each C_i is continuous, this lemma implies $C_i(\overline{\psi}_{n,I^*}, \overline{\psi}_{n,i}) \longrightarrow C_i(\beta, \psi^\beta_i)$ for all $i \neq I^*$. Proposition 6 then follows by invoking Proposition 5, which establishes the optimality of the allocation $\boldsymbol{\psi}^\beta$.

The proof is broken into a number of steps. The first step shows that the posterior is consistent under any of the proposed algorithms. As noted in step 2, this immediately implies that $\alpha_{n,I^*} \to 1$, and that each algorithm allocates β–fraction of measurement effort to I^* in the long run. The main results are left to steps 3-5.

Step 3 shows that whenever an algorithm allocates too much measurement effort to some design $i \neq I^*$, in the sense that $\overline{\psi}_{n,i} > \psi^\beta_i + \delta$ for a constant $\delta > 0$, then $\alpha_{n,i}$ will be exponentially small compared $\max_{j \neq I^*} \alpha_{n,j}$. Step 4 observes that as a result, once n is sufficiently large, none of the proposed algorithms ever sample a design with $\overline{\psi}_{n,i} > \psi^\beta_i + \delta$. The proof is completed in step 5, by noting that since none the proposed algorithms can allocate a fraction of effort strictly greater than ψ^β_i to design i, they must allocate exactly the effort level ψ^β_i to each design i in the long run.

Step 1: *Let $\tilde{\Theta} \subset \Theta$ be an open set containing $\boldsymbol{\theta}^*$. Under either PTS, TTPPS, or TTVS, with parameter $\beta > 0$*

$$\Pi_n(\tilde{\Theta}) \longrightarrow 1.$$

Proof. By Proposition 7 established in Section B.2 , it's enough to show that each design is measured infinitely often. Therefore, we would like to show that $S_{n,i} \to \infty$ almost surely, or equivalently (see Section A, fact 2) that $\sum_{\ell=1}^n \psi_{\ell,i} \to \infty$.

By contradiction, suppose designs in $\mathcal{I} \subset \{1, ..., k\}$ are sampled only a finite number of times. By Proposition 7, for any $\epsilon > 0$

$$\Pi_n\left(\{\boldsymbol{\theta} : \theta_i \in (\theta^*_i - \epsilon, \theta^*_i + \epsilon) \longrightarrow 1 \qquad \forall i \notin \mathcal{I}. \right. \tag{22}$$

Therefore, if $\rho^* = \max_{i \notin \mathcal{I}} \theta^*_i$ is the quality of the best design among those that are sampled infinitely often,

$$\Pi_n\left(\{\boldsymbol{\theta} : \max_{i \notin \mathcal{I}} \theta_i \geq \rho^* + \epsilon\}\right) \to 0.$$

However, as long as ϵ is chosen to be small enough that $\rho^* + 2\epsilon < \overline{\theta}$, by the second part of Proposition 7,

$$\liminf_{n\to\infty} \Pi_n\left(\{\boldsymbol{\theta} : \theta_i > \rho^* + 2\epsilon\}\right) > 0 \qquad \forall i \in \mathcal{I}.$$

As a result

$$\liminf_{n\to\infty} \Pi_n\left(\{\boldsymbol{\theta} : \max_{i \in \mathcal{I}} \theta_i > \max_{i \notin \mathcal{I}} \theta_i + \epsilon\}\right) > 0 \tag{23}$$

and therefore

$$\liminf_{n\to\infty} \sum_{i \in \mathcal{I}} \alpha_{n,i} > 0. \implies \exists i \in \mathcal{I} \quad s.t. \quad \liminf_{n\to\infty} \alpha_{n,i} > 0. \tag{24}$$

Analogously, (23) implies

$$\exists i \in \mathcal{I} \quad s.t. \quad \liminf_{n\to\infty} V_{n,i} > 0. \tag{25}$$

In words, the quality of designs that are sampled an infinite number of times is identified to arbitrary precision. But for designs that are only measured a finite number of times, there is always nonzero probability under the posterior that one of them significantly exceeds the highest quality that has been confidently identified. Therefore, the posterior probability that some design in $\mathcal{I}$ is optimal cannot tend to zero. We will show that this implies the PTS, TTPS, and TTVS rules will measure designs in the set $\mathcal{I}$ an infinite number of times, yielding a contradiction.

Proof for PTS: Under the PTS sampling rule, any action in i is sampled with probability at least $\psi_{n,i} \geq \beta\alpha_{n,i}$. Therefore, by (24), there must be some design $i \in \mathcal{I}$ with $\sum_{n\in\mathbb{N}} \psi_{n,i} = \infty$. As stated at the beginning of the proof, this implies $S_{n,i} \to \infty$ and yields a contradiction.

Proof for TTPS: We show that there is a finite time after which one of the top two designs must be in the set $\mathcal{I}$ for all remaining periods. This immediately implies that the set $\mathcal{I}$ will be sampled infinitely often by the nature of the TTPS algorithm. By hypothesis, there exists a finite time τ after which no design in $\mathcal{I}$ is sampled. By (24), there is a time $\tau' \geq \tau$ and some probability $\alpha' > 0$ such that $\max_{i\in\mathcal{I}} \alpha_{n,i} > \alpha'$ for all $n \geq \tau'$. 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 (22), the algorithm identifies $\arg\max_{i\notin\mathcal{I}} \theta^*_i$ with certainty, and $\alpha_{n,i} \to 0$ for every $i \notin \mathcal{I}$ except for I. Then there is a time $\tau'' > \tau'$ such that for $n \geq \tau''$

$$\max_{i\in\mathcal{I}} \alpha_{n,i} > \alpha' \qquad \max_{i\notin\mathcal{I}, i\neq I} \alpha_{n,i} < \alpha'.$$

When this occurs at least one of the top two designs (those with highest value $\alpha_{n,i}$) must be in the set $\mathcal{I}$, which implies designs in that set are measured infinitely often, yielding a contradiction.

Proof for TTVS: The proof is essentially identical to that for TTPS. By (25) there is a time $\tau' \geq \tau$ and some number $v' > 0$ such that $\max_{i\in\mathcal{I}} V_{n,i} > v'$ for all $n \geq \tau'$. As before, if $I = \arg\max_{i\notin\mathcal{I}} \theta^*_i$ then $\alpha_{n,i} \to 0$ for every $i \in \mathcal{I}$ except for I. Since $V_{n,i} \leq (u(\overline{\theta}) - u(\underline{\theta}))\alpha_{n,u}$, this implies $V_{n,i} \to 0$ for every $i \in \mathcal{I}$ except for I. Then there is a time $\tau'' > \tau'$ such that for $n \geq \tau''$

$$\max_{i\in\mathcal{I}} V_{n,i} > v' \qquad \max_{i\notin\mathcal{I}, i\neq I} V_{n,i} < v'.$$

As before, this implies designs in the set $\mathcal{I}$ are sampled infinitely often. □

The next step follows as a corollary.

Step 2: *Under TTPS, TTTVS, or TPS with parameter β, $\overline{\psi}_{n,I^*} \to \beta$ as $n \to \infty$. In addition, $\alpha_{n,I^*} \to 1$, $\max_{i\neq I^*} V_{n,i} \to 0$ and $V_{n,I^*} \to v_{I^*}(\boldsymbol{\theta}^*) > 0$.*

Proof sketch. Since there is an open set around $\boldsymbol{\theta}^*$ that does not intersect Θ_{-I^*}., this lemma implies that $\Pi_n(\Theta_{-I^*}) \to 0$ under any of the proposed algorithms. Equivalently, it implies that $\alpha_{n,I^*} \to 1$. This immediately implies that $\overline{\alpha}_{n,I^*} \to \beta$ for under PTS and TTPS. For PTS, this implies fraction of times where the first sample drawn from $\boldsymbol{\alpha}_n$ is I^* tends to 1, and therefore the average effort allocated to measuring I^* tends to β.For Top-Two Probability sampling, there is a finite time after which the top design is always I^* (i.e. $|\{n : \arg\max_i \alpha_{n,i} \neq I^*\}| < \infty$), and so the same property holds. Turning to TTVS, for any $i \neq I^*$, $V_{n,i} \to 0$ since $\alpha_{n,i} \to 0$. The claim that $V_{n,I^*} \to v_{I^*}(\boldsymbol{\theta}^*)$ follows from step 1, and the continuity of $v_i(\boldsymbol{\theta})$. Then, as with TTPS, for Top Two Value sampling

there is a finite time after which the top-design is I^* for all subsequent periods, which implies $\overline{\psi}_{n,I^*} \to \beta$. □

Step 3: *Fix any $\delta > 0$ and $j \neq I^*$. Under any allocation rule, if $\overline{\psi}_{n,I^*} \to \beta$, then there exists $\delta' > 0$ and a sequence ϵ_n with $\epsilon_n \to 0$ such that for any $n \in \mathbb{N}$,*

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

Proof. Since $\Pi_n(\Theta_{-I^*}) = \sum_{i \neq I^*} \alpha_{n,i}$, $\Pi_n(\Theta_{-I^*}) \doteq \max_{i \neq I^*} \alpha_{n,i}$. Then, by invoking Proposition (5), since $\overline{\boldsymbol{\psi}}_{n,I^*} \longrightarrow \beta$,

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

Now, by Proposition 3 and Lemma 1,

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

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

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

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

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

□

The next step shows that, for sufficiently large n, none of the proposed algorithms ever measures a design $j \neq I^*$ to which too much measurement effort has been allocated, in the sense that $\overline{\psi}_{n,j} \geq \psi_j^\beta + \delta$. This is shown separately for each algorithm.

Step 4: Under TTPS, TTVS, or PTS with parameter β, for any $j \neq I^*$, and $\delta > 0$,

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

Proof for TTPS. Let $\hat{I}_n = \arg\max_i \alpha_{n,i}$ and $\hat{J}_n = \arg\max_{i \neq \hat{I}_n} \alpha_{n,i}$. Top two sampling always measures one of these two designs. By Step 2, $\alpha_{n,I^*} \to 1$, so there exists a finite time τ such that $\hat{I}_n = I^*$ for all $n \geq \tau$. Therefore, By Step 3, we can choose $\tau'' \geq \tau'$ such that for all $n \geq \tau''$

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

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

Proof for TTVS. The proof is essentially identical to that for TTPS. Let $\hat{I}_n = \arg\max_i V_{n,i}$ and $\hat{J}_n = \arg\max_{i \neq \hat{I}_n} V_{n,i}$. TTVS always measures one of these two designs. By Step 2, $V_{n,I^*} \to v_{I^*}(\boldsymbol{\theta}^*) > 0$ and $\max_{i \neq I^*} V_{n,i} \to 0$, so there exists a finite time τ such that $\hat{I}_n = I^*$ for all $n \geq \tau$. By Lemma 3, $\alpha_{n,i} \doteq V_{n,i}$, and so, applying step 3, we can choose $\tau'' \geq \tau'$ such that for all $n \geq \tau''$

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

Again, this shows there is a finite time τ' after which TTVS never allocates any measurement effort to design j when $\overline{\psi}_{n,j} \geq \psi_j^\beta + \delta$. □

Proof for PTS. As before, let $\hat{I}_n = \arg\max_i \alpha_{n,i}$, and $\hat{J}_n = \arg\max_{i \neq \hat{I}_n} \alpha_{n,i}$. As before, for each sample path there is a finite time $\tau < \infty$ such that for all $n \geq \tau$, $\hat{I}_n = I^*$ and therefore $\hat{J}_n = \arg\max_{i \neq I^*} \alpha_{n,i}$. Recall the definition of the PTS algorithm. It draws at each time step a random design X_n where the probability $X_n = i$ is $\alpha_{n,i}$ Then with probability β, it chooses X_n. With probability $1 - \beta$ it chooses S'_n where the probability $S'_n = i$ is $\alpha_{n,i} / \sum_{\ell \neq X_n} \alpha_{n,\ell}$. Note that $\sum_{\ell \neq X_n} \alpha_{n,\ell} \geq \alpha_{n,\hat{J}_n}$, which implies the probability of playing action j at time n is bounded as

$$\psi_{n,j} \leq \beta \alpha_{n,j} + (1 - \beta) \frac{\alpha_{n,j}}{\alpha_{n,J_n}} \leq \frac{\alpha_{n,j}}{\alpha_{n,J_n}}.$$

For $n \geq \tau$, this means $\psi_{n,j} \leq \alpha_{n,j} / (\max_{i \neq I^*} \alpha_{n,i})$. By Step 3, there is a constant $\delta' > 0$ and a sequence $\epsilon_n \to 0$ such that

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

Therefore

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

□

Step 5: Consider any adaptive allocation rule. If $\overline{\psi}_{n,I^*} \to \beta$ and

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

then $\overline{\boldsymbol{\psi}}_n \to \boldsymbol{\psi}^\beta$.

Proof. Fix some $j \neq I^*$. We first show $\liminf_{n \to \infty} \overline{\psi}_{n,j} \leq \psi_j^*$. Suppose otherwise. Then, with positive probability, for some $\delta > 0$, there exists N such that for all $n \geq N$, $\overline{\psi}_{n,j} \geq \psi_j^* + \delta$. But then,

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

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

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

$$\limsup_{n \to \infty} \overline{\psi}_{n,j} > \psi_j^\beta \qquad \& \qquad \liminf_{n \to \infty} \overline{\psi}_{n,j} \leq \psi_j^\beta.$$

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

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

which violates the hypothesis.

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