On Multi-Armed Bandit Designs for Dose-Finding Trials

Maryam Aziz, Emilie Kaufmann, Marie-Karelle Riviere

On Multi-Armed Bandit Designs for Dose-Finding Trials

Maryam Aziz[1], Emilie Kaufmann[2] and Marie-Karelle Riviere[3]

[1]Spotify, [2]CNRS & Univ. Lille, CRIStAL (UMR 9189), Inria SequeL
[3] Statistical Methodology Group, Biostatistics and Programming department, Sanofi R&D

Abstract

We study the problem of finding the optimal dosage in early stage clinical trials through the multi-armed bandit lens. We advocate the use of the Thompson Sampling principle, a flexible algorithm that can accommodate different types of monotonicity assumptions on the toxicity and efficacy of the doses. For the simplest version of Thompson Sampling, based on a uniform prior distribution for each dose, we provide finite-time upper bounds on the number of sub-optimal dose selections, which is unprecedented for dose-finding algorithms. Through a large simulation study, we then show that variants of Thompson Sampling based on more sophisticated prior distributions outperform state-of-the-art dose identification algorithms in different types of dose-finding studies that occur in phase I or phase I/II trials.

1 Introduction

Multi-armed bandit models were originally introduced in the 1930's as a simple model for a (phase III) clinical trial in which one control treatment is tried against one alternative (Thompson, 1933). While those models are nowadays widely studied with completely different applications in mind, like online advertisement (Chapelle and Li, 2011), recommender systems (Li et al., 2010) or cognitive radios (Anandkumar et al., 2011), there has been a surge of interest in the use of bandit algorithms for clinical trials (see Villar et al. (2015)). More broadly, Adaptive Clinical Trials have received an increased attention (Pallmann et al., 2018) as the Food and Drug Administration recently updated a draft of guidelines for their actual use (Food and Drugs Administration (FDA), 2018). In this paper, we focus on adaptive designs for phase I and phase I/II clinical trials for single-agent in oncology, for which adaptations of the original bandit algorithms may be of interest.

Phase I trials are the first stage of testing in human subjects. Their goal is to evaluate the safety (and feasibility) of the treatment and identify its side effects. For non-life-threatening diseases, phase I trials are usually conducted on human volunteers. In life-threatening diseases such as cancer or AIDS, phase I studies are conducted with patients because of the aggressiveness and possible harmfulness of the treatments, possible systemic treatment effects, and the high interest in the new drug's efficacy in those patients directly. The aim of a phase I dose-finding study is to *determine the most appropriate dose level that should be used in further phases of the clinical trials*. Traditionally, the focus is on determining the highest dose with acceptable toxicity called the Maximum Tolerated Dose (MTD). Once the initial safety of the drug has been confirmed in phase I trials, phase II trials are performed on larger groups and are designed to establish the efficacy of the drug and confirm the safety identified in phase I. In phase II dose-finding studies, the dose-efficacy relationship is modeled in order to estimate the smallest dose to obtain a desired efficacy, called the minimal effective dose (MED). Approaches that use both efficacy and toxicity to find an optimal dose are called phase I/II designs. If the new potential treatment shows some efficacy in phase II, it is compared to alternative

treatments in phase III. We here consider two classes of algorithms for dose-finding in early stage trials: algorithms which consider only toxicity, suited for phase I trials, and algorithms which consider both toxicity and efficacy, suited for phase I/II trials.

Until recently, cytotoxic agents were the main agent of anti-tumor drug development. A common assumption for these agents is that both toxicity and efficacy of the treatment are monotonically increasing with the dose (Chevret, 2006). Hence, only toxicity is required to determine the optimal dose which is then the Maximum Tolerated Dose. From a statistical perspective, the MTD is often defined as the dose level closest to an acceptable targeted toxicity probability fixed prior to the trial onset (Faries, 1994; Storer, 1989). However, Molecularly Targeted Agents (MTAs) have emerged as a new treatment option in oncology that have changed the practice of cancer patient care (Postel-Vinay et al., 2009; Le Tourneau et al., 2010, 2011, 2012). Previously-common assumptions do not necessarily hold for MTAs. Although toxicity is still assumed to be increasing with the dose, it may be so low that the trial cannot be driven by toxicity occurrence only. Efficacy needs to be studied jointly with toxicity, so that the most appropriate dose is not just the MTD. In particular, for some mechanisms of action, a plateau of efficacy can be observed when increasing the dose (Hoering et al., 2011), for instance when the targeted receptors are saturated. In this paper, we aim at providing a unified approach that can be used both for phase I trials involving cytotoxic agents and phase I/II trials involving MTAs.

Phase I cytotoxic clinical trials in oncology involve several ethical concerns. Therefore, in order to gather information about the dose-toxicity relationship it is not possible to include a large number of patients and randomize them at each different dose level considered in the trial. Patients treated with dose levels over the MTD would be exposed to very high toxicity, and patients treated at low dose levels would be administrated ineffective dose levels. In addition, the total sample size is often very limited. For these reasons, the doses to be allocated should be selected sequentially, taking into account the outcomes of the previous allocated doses, with ideally two objectives in mind: finding the MTD (which is crucial for the next stages of the trial) and treating as many trial participants as possible with this MTD. This trade-off between treatment (curing patients during the study) and experimentation (finding the best treatment) is a common issue in clinical trials. By viewing optimal dose identification as a particular multi-armed bandit problem, this trade-off can be rephrased as a trade-off between rewards and error probability, two performance measures that are well-studied in the bandit literature and that are known to be somewhat antagonistic (see Bubeck et al. (2011); Kaufmann and Garivier (2017)).

In this paper, we investigate the use of Thompson Sampling (Thompson, 1933) for dose-finding clinical trials. This Bayesian algorithm has gained a lot of popularity in the machine learning community for its successful use for reward maximization in bandit models (see, e.g., Chapelle and Li (2011)). Interestingly, in the growing literature on Bayesian Adaptive Designs (Berry, 2006; Berry et al., 2010), several designs that may be viewed as variants of Thompson Sampling have been proposed for other types of clinical trials in which different treatments are compared (Thall and Wathen, 2007; Satlin et al., 2016). However, to the best of our knowledge, the use of Thompson Sampling has not been investigated yet for dose-finding trials, and the present paper aims to fill this gap. We show that, unlike other bandit algorithms that are better suited for phase III trials, Thompson Sampling can indeed be naturally adapted to dose-finding trials.

Our first contribution is a theoretical study in the context of MTD identification showing that the simplest version of Thompson Sampling based on independent prior distributions for each arm asymptotically minimizes the number of sub-optimal allocations during the trial. Albeit asymptotic, this sanity-check for Thompson Sampling with a simple prior motivates our investigation for its use with more realistic prior distributions, where theoretical guarantees are harder to obtain. Our second contribution is to show that Thompson Sampling using more sophisticated prior distributions can compete with state-of-the art dose-finding algorithms.

We indeed show that the algorithm can exploit the monotonicity assumption on the toxicity probabilities that are common for MTD identification (Section 4.1), but also deal with more complex assumptions on both the toxicity and efficacy probabilities that are relevant for trials involving MTAs (Section 4.2). Through extensive experiments on simulated clinical trials we show that our Thompson Sampling variants typically outperform state-of-the-art dose-finding algorithms. Finally, we propose a discussion revisiting the treatment versus experimentation trade-off through a bandit lens, and explain why an adaptation of existing best arm identification designs (Audibert et al., 2010; Karnin et al., 2013) seems currently less promising for dose-finding clinical trials.

The paper is structured as follows. In Section 2, we present a multi-armed bandit (MAB) model for the MTD identification problem and introduce the Thompson Sampling algorithm. In Section 3, we propose an analysis of Thompson Sampling with independent Beta priors on the toxicity of each dose: We provide finite-time upper-bounds on the number of sub-optimal selections, which match an (asymptotic) lower bound on those quantities. Then in Section 4, we show that Thompson Sampling can leverage the usual monotonicity assumptions in dose-finding clinical trials. In Section 5, we report the results of a large simulation study to assess the quality of the proposed design. Finally in Section 6, we propose a discussion on the use of alternative bandit methods.

2 Maximum Tolerated Dose Identification as a Bandit Problem

In this section, we propose a simple statistical model for the MTD identification problem in phase I clinical trials and show that it can be viewed as a particular multi-armed bandit problem.

A dose-finding study involves a number K of dose levels that have been chosen by physicians based on preliminary experiments (K is usually a number between 3 and 10). Denoting by p_k the (unknown) toxicity probability of dose k, the Maximum Tolerated Dose (MTD) is defined as the dose with a toxicity probability closest to a target:

$$k^* \in \underset{k\in\{1,\dots,K\}}{\operatorname{argmin}} |\theta - p_k|,$$

where θ is the pre-specified targeted toxicity probability (typically between 0.2 and 0.35). For clinical trials in life-threatening diseases, efficacy is often assumed to be increasing with toxicity, hence the MTD is the most appropriate dose to further investigate in the rest of the trial. However, we shall see in Section 4 that under different assumptions the optimal dose may be defined differently.

2.1 A (Bandit) Model for MTD Identification

A MTD identification algorithm proceeds sequentially: at round t a dose $D_t \in \{1,\dots,K\}$ is selected and administered to a patient for whom a toxicity response is observed. A binary outcome X_t is revealed where $X_t = 1$ indicates that a harmful side-effect occurred and $X_t = 0$ indicates than no harmful side-effect occurred. We assume that X_t is drawn from a Bernoulli distribution with mean p_{D_t} and is independent from previous observations. The *selection rule* for choosing the next dose level to be administered is sequential in that it uses the past toxicity observations to determine the dose to administer to the next patient. More formally, D_t is $\mathcal{F}_{t-1}$-measurable where $\mathcal{F}_t = \sigma(U_0, D_1, X_1, U_1, \dots, D_t, X_t, U_t)$ is the σ-field generated by the observations made with the first t patients and the possible exogenous randomness used in each round t, $U_{t-1} \sim \mathcal{U}([0,1])$. Along with this selection rule, a ($\mathcal{F}_t$-measurable) *recommendation rule* $\hat{k}_t$ indicates which dose would be recommended as the MTD, if the experiments were to be stopped after t patients.

Usually in clinical trials the total number of patients n is fixed in advance and the first objective is to ensure that the dose $\hat{k}_n$ recommended at the end of the trial is close to the MTD, k^*, but there is also an incentive to treat as many patients as possible with the MTD during the trial. Letting $N_k(t) = \sum_{s=1}^{t} \mathbb{1}_{(D_s=k)}$ be the number of time dose k has been given to one of the first t patients, this second objective can be formalized as that of minimizing $N_k(n)$ for $k \neq k^*$. In the clinical trial literature, empirical evaluations of dose-finding designs usually report both the empirical distribution of the recommendation strategy $\hat{k}_n$ (that should be concentrated on the MTD) and estimates of $\mathbb{E}[N_k(n)]/n$ for all doses k to assess the quality of the selection strategy in terms of allocating MTD as often as possible.

The sequential interaction protocol described above is reminiscent of a stochastic multi-armed bandit (MAB) problem (see Lattimore and Szepesvari (2018) for a recent survey). A MAB model refers to a situation in which an agent sequentially chooses arms (here doses) and gets to observe a realization of an underlying probability distribution (here a Bernoulli distribution with mean being the probability that the chosen dose is toxic). Different objectives have been considered in the bandit literature, but most of them are related to *learning the arm with largest mean*, whereas in the context of clinical trials we are rather concerned with the arm which is the closest to some threshold.

2.2 Thompson Sampling for MTD Identification

Early works on bandit models (Robbins, 1952; Lai and Robbins, 1985) mostly consider a *reward maximization* objective: The samples (X_t) are viewed as rewards, and the goal is to maximize the sum of these rewards, which boils down to choosing the arm with largest mean as often as possible. This problem was originally introduced in the 1930s in the context of phase III clinical trials (Thompson, 1933). In this context, each arm models the response to a particular treatment, and maximizing rewards amounts to giving the treatment with largest probability of success to as many patients as possible. This suggests a phase III trial is designed for treating as many patients as possible with the best treatment rather than identifying it. The trade-off between treatment and identification is also relevant for MTD identification: besides finding the MTD another objective is to treat as many patients as possible with it during the trial.

Reward maximization in a Bernoulli bandit model is a well-studied problem (Jacko, 2019). In particular, it is known since (Lai and Robbins, 1985) that any algorithm that performs well on every bandit instance should select each sub-optimal arm k more than $C_k \log(n)$ times, where C_k is some constant, in a regime of large values of n. Algorithms with finite-time upper bounds on the number of sub-optimal selections have been exhibited (Auer et al., 2002; Audibert et al., 2009), some of which match the aforementioned lower bound on the number of sub-optimal selections (Cappé et al., 2013). In the context of MTD identification, we are also concerned about *minimizing the number of sub-optimal selections* but with a different notion of optimal arm: the MTD instead of the arm with largest mean.

Algorithms for maximizing rewards in a bandit model mostly fall in two categories: frequentist algorithms, based on upper-confidence bounds (UCB) for the unknown means of the arms (popularized by Katehakis and Robbins (1995); Auer et al. (2002)) and Bayesian algorithms, that exploit a posterior distribution on the means (see, e.g. Powell and Ryzhov (2012); Kaufmann et al. (2012a)). Among those, Thompson Sampling (TS) is a popular approach, known for its practical successes beyond simple bandit problems (Agrawal and Goyal, 2013b; Agrawal and Jia, 2017). In the context of clinical trials, variants of Thompson Sampling have been notably studied for phase III clinical trials involving two treatments (see Thall and Wathen (2007) and references therein), or for adaptive trials involving interim analyses (Satlin et al., 2016). Strong theoretical properties have also been established for this algorithm in simple models. In particular, Thompson Sampling was proved to be asymptotically optimal for Bernoulli bandit models (Kaufmann et al., 2012b; Agrawal and Goyal, 2013a).

Thompson Sampling, also known as probability matching, implements the following simple Bayesian heuristic. Given a prior distribution over the arms, at each round an arm is selected at random according to its posterior probability of being optimal. In this paper, we advocate the use of Thompson Sampling for dose-finding, using the appropriate notion of optimality. In particular, Thompson Sampling for MTD identification consists of selecting a dose at random according to its posterior probability of being the MTD. Given a prior distribution Π^0 on the vector of toxicity probabilities, $\boldsymbol{p} = (p_1, \dots, p_K) \in [0,1]^K$, a posterior distribution Π^t can be computed by taking into account the first t observations. A possible implementation of Thompson Sampling consists of drawing a sample $\boldsymbol{\theta}(t) = (\theta_1(t), \dots, \theta_K(t))$ from the posterior distribution Π^t and selecting at round $t+1$ the dose that is the MTD in the sampled model: $D_{t+1} = \text{argmin}_k \, |\theta_k(t) - \theta|$. There are several possible choices for the recommendation rule $\hat{k}_t$, which are discussed in the upcoming sections.

2.3 Why Thompson Sampling?

Thompson Sampling is by far not the only existing bandit algorithm, yet other algorithms may not be as easily adaptable to the MTD identification problem, which justifies our focus on this algorithm.

Indeed, Thompson Sampling only requires defining some notion of *optimal arm* (or arm to discover), which is naturally defined as the arm with mean closest to the threshold θ in the MTD identification problem. Many other popular bandit algorithms instead require a *value* to be assigned to each sampled arm, and require the optimal arm to be the arm with largest expected value. This is the case for the frequentist *optimistic* (UCB) algorithms (see, e.g., Auer et al. (2002); Cappé et al. (2013)), which construct confidence intervals on the expected value of each arm and select the arm which has the largest statistically plausible expected value (i.e. the largest Upper Confidence Bound). Adapting this optimism in face of uncertainty principle for MTD identification is not straightforward: one can certainly build confidence intervals on the toxicity probability of each dose (several of them may contain the MTD), but there is no natural way to define a "best plausible value" for each dose in that case.

In the literature on Bayesian ranking and selection, value-based approaches have also been proposed. Some algorithms are indeed based on defining some Expected Value of Information (Chick, 2006). Among those, knowledge gradient methods (Powell and Ryzhov, 2012) are particularly interesting since they permit handling correlations between arms. For example Xie et al. (2016) consider a prior distribution over the arms' means which is a multivariate Gaussian, and Wang et al. (2016) consider a Bayesian logistic model (where a Laplace approximation is used for Bayesian inference). However, the proposed algorithms are both tailored to finding an arm a maximizing $\mathbb{E}[V(a, D)]$ for some function V that depends on a random variable D under which the expectation is taken (like other algorithms from the Bayesian Optimization (BO) literature (Brochu et al., 2010)). The MTD identification problem cannot naturally be cast in this framework, and adapting, e.g., knowledge gradient methods would require defining an appropriate notion of value of information in this setting. This is why we focused on a Bayesian approach which is easier to adapt to MTD identification, Thompson Sampling.

3 Independent Thompson Sampling: an Asymptotically Optimal Algorithm

Inspired by the bandit literature, we introduce the simplest version of Thompson Sampling, that assumes independent uniform prior distributions on the probability of toxicity of each dose. We refer to this algorithm as Independent Thompson Sampling and propose some theoretical guarantees for it.

3.1 Algorithm Description

The prior distribution on $\boldsymbol{p} = (p_1, \dots, p_K)$ is $\Pi^0 = \bigotimes_{i=1}^K \pi_k^0$, where $\pi_k^0 = \mathcal{U}([0,1])$ is a uniform distribution. Letting π_k^t be the posterior distribution of p_k given the observations from the first t patients, the posterior distribution also has a product form, $\Pi^t = \bigotimes_{i=1}^K \pi_k^t$. Moreover, each π_k^t can be made explicit: π_k^t is a $\text{Beta}(S_k(t)+1, N_k(t) - S_k(t) + 1)$ distribution where $S_k(t) = \sum_{s=1}^t X_s \mathbb{1}_{(D_s=k)}$ is the sum of rewards obtained from arm k and D_s is the dose allocated at time s.

The selection rule of Independent Thompson Sampling is simple: a sample from the posterior distribution on the toxicity probability of each dose is generated, and the dose for which the sample is closest to the threshold is selected:

$$\begin{cases} \forall k \in \{1, K\},\ \theta_k(t) \sim \pi_k^t \\ D_{t+1} = \text{argmin}_k\, |\theta_k(t) - \theta|. \end{cases}$$

Several recommendation rules may be used for Independent Thompson Sampling. As the randomization induces some exploration, recommending $\hat{k}_t = D_{t+1}$ is not a good idea. Inspired by what is proposed by Bubeck et al. (2011) for assigning a recommendation rule to rewards maximizing algorithms, a first idea is to recommend $\hat{k}_t = \text{argmin}_k\, |\hat{\mu}_k(t) - \theta|$, where $\hat{\mu}_k(t)$ is the empirical mean of dose k after the t-th patient of the study. Leveraging the fact that TS is supposed to allocate the MTD most of the time, we could also select $\hat{k}_t = \text{argmax}_k\, N_k(t)$ or pick $\hat{k}_t$ uniformly at random among the allocated doses.

3.2 Upper Bound on the Number of Sub-Optimal Selections

For the classical rewards maximization problem, the first finite-time analysis of Thompson Sampling for Bernoulli bandits dates back to Agrawal and Goyal (2012) and was further improved by Kaufmann et al. (2012b); Agrawal and Goyal (2013a). In Appendix A, building on the analysis of Agrawal and Goyal (2013a), we prove the following for Thompson Sampling applied to MTD identification.

Theorem 1. *Introducing for every $k \neq k^*$ the quantity*

$$d_k^* := \underset{d \in \{p_{k^*}, 2\theta - p_{k^*}\}}{\text{argmin}} |p_k - d|,$$

Independent Thompson Sampling satisfies the following. For all $\varepsilon > 0$, there exists a constant $C_{\varepsilon,\theta,\boldsymbol{p}}$ (depending on ε, the threshold θ and the toxicity probabilities) such that for all k : $|p_k - \theta| \neq |\theta - p_{k^*}|$,

$$\mathbb{E}[N_k(n)] \leq \frac{1+\varepsilon}{\text{kl}(p_k, d_k^*)} \log(n) + C_{\varepsilon,\theta,\boldsymbol{p}},$$

where $\text{kl}(x,y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$ is the binary Kullback-Leibler divergence.

Theorem 1 shows that the total number of allocations to a sub-optimal dose in a trial involving n patients is logarithmic in n, which justifies that the MTD is given most of the time, at least in a regime of large values of n (as the second order term can be large). Also, this bounds tells us that in this regime each sub-optimal dose is allocated in inverse proportion of $\text{kl}(p_k, d_k^*)$, which can be seen as a distance between dose k and an optimal dose with toxicity probability d_k^* which is illustrated in Figure 1.

The lower bound given in Theorem 2 below furthermore shows that Independent Thompson Sampling actually achieves the *minimal number of sub-optimal allocations* when n grows large.

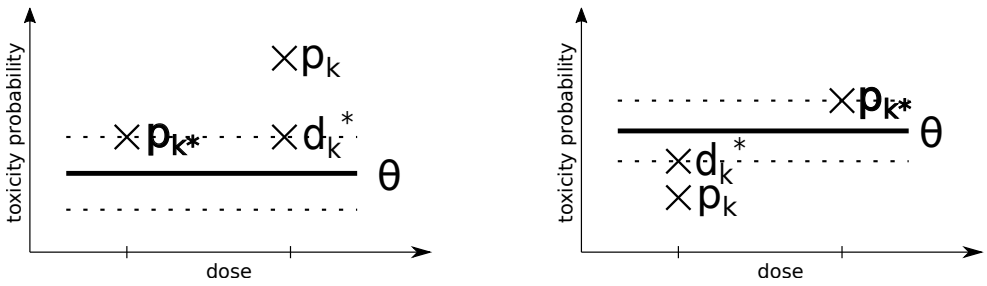


Figure 1: Optimal dose d_k^* associated with dose k. In some cases $d_k^* = p_{k^*}$ (left), in others $d_k^* = 2\theta - p_{k^*}$ (right), which is symmetric to the MTD with respect to threshold θ.

Theorem 2. *We define a uniformly efficient design as a design satisfying for all possible toxicity probabilities* $\boldsymbol{p}$*, for all* $\alpha \in]0,1[$*, for all* $k : |\theta - p_k| \neq |\theta - p_{k^*}|$*,* $\mathbb{E}[N_k(n)] = o(n^\alpha)$ *when* n *goes to infinity. If* $p_{k^*} \neq \theta$*, any uniformly efficient design satisfies, for all* k*:* $|\theta - p_k| \neq |\theta - p_{k^*}|$*,*

$$\liminf_{n\to\infty} \frac{\mathbb{E}[N_k(n)]}{\log(n)} \geq \frac{1}{\mathrm{kl}(p_k, d_k^*)}.$$

Theorem 2 can be viewed as a counterpart of the Lai and Robbins lower bound for classical bandits (Lai and Robbins, 1985) and can be easily derived using recent change-of-measure tools (see Garivier et al. (2019b)). Its proof is given in Appendix B for the sake of completeness.

3.3 Upper Bound on the Error Probability

If the recommendation rule $\hat{k}_n$ consists of selecting uniformly at random a dose among the doses that were allocated during the trial, $\{D_1, \ldots, D_n\}$, it follows from Theorem 1 that

$$\mathbb{P}\left(\hat{k}_n \neq k^*\right) = \sum_{k \neq k^*} \frac{\mathbb{E}[N_k(n)]}{n} \leq \frac{D \ln(n)}{n}, \tag{1}$$

where D is a (possibly large) problem-dependent constant. Hence finite-time upper bounds on the number of sub-optimal selection lead to *non-asymptotic upper bound on the error probability* of the design. Note that for the state-of-the-art dose-finding designs it is not known whether such results can be obtained; the only results available provide conditions for *consistency*. For example Shen and O'Quigley (1996); Cheung and Chappell (2002) exhibit some conditions on the toxicity probabilities under which a classical design called the CRM is such that $\hat{k}_n$ converges almost surely to k^*.

This being said, the upper bound (1) is not very informative, as a very large number of patients is needed for the upper bound to be at least smaller than 1, and one could expect to have an upper bound that is exponentially decreasing with n. As we shall see in Section 6, an adaptation of a best arm identification algorithm (Karnin et al., 2013) leads to such an upper bound, but may be less desirable for clinical trials from an ethical point of view. This is why we rather chose to investigate in what follows several variants of Thompson Sampling coupled with an appropriate recommendation rule.

By using uniform and independent priors on each toxicity probability, Independent Thompson Sampling is the simplest possible implementation of Thompson Sampling. We now explain that using a more sophisticated prior distribution allows the algorithm to leverage some particular constraints of the dose-finding problem, like increasing toxicities or a plateau of efficacy.

4 Exploiting Monotonicity Constraints with Thompson Sampling

Independent Thompson Sampling is an adaptation of a state-of-the-art bandit algorithm for identifying the MTD that does not leverage any prior knowledge on (e.g.) the ordering of the arms' means. While it can be argued that when testing drug combinations no natural ordering between the doses exists (see, e.g., Mozgunov and Jaki (2017)), in most cases monotonicity assumptions can speed up learning.

A typical assumption in phase I studies is that both efficacy and toxicity are increasing with the dose. We show in Section 4.1 that Thompson Sampling using an appropriate prior is competitive with state-of-the-art phase I approaches leveraging the monotonicity. In Section 4.2, we further show that Thompson Sampling is a flexible method that can be useful in phase I/II trials, under more complex monotonicity assumptions on both toxicity an efficacy. More specifically, we show it can handle an efficacy "plateau," where efficacy may be non-increasing after a certain dose level.

4.1 Thompson Sampling for Increasing Toxicities: A Phase I Design

In a phase I study in which both toxicity and efficacy are increasing with the dose, the MTD is the most relevant dose to allocate in further stages. We now focus on algorithms leveraging the extra information that $p_1 \leq \cdots \leq p_k$. To exploit this structure, *escalation procedures* have been developed in the literature, the most famous being the "3+3" design (Storer, 1989). In this design, adjusted for $\theta = 0.33$, the lowest dose is first given to 3 patients. If no patient experiences toxic effects, one escalates to the next dose and repeats the process. If one patient experiences toxicity, the dose is given to 3 more patients, and if less than two patients among the 6 experience toxicity, one escalates to the next dose. Otherwise the trial is stopped, which is also the case if from the beginning 2 out of the 3 patients experience a toxic effect. Upon stopping, the previous dose is recommended as the MTD, or all doses are deemed too toxic if one stops at the first dose level. Although it is clear that the guarantees in terms of error probability (or sub-optimal selections) are very weak, "3+3" is still often used in practice.

Alternative to this first design are variants of the Continuous Reassessment Method (CRM), proposed by O'Quigley et al. (1990). The CRM uses a Bayesian model that combines a parametric dose/toxicity relationship with a prior on the model parameters. Under this model, CRM appears as a greedy strategy that selects in each round the dose whose expected toxicity under the posterior distribution is closest to the threshold. We propose in this section several variants of Thompson Sampling based on the same Bayesian model, but that favor (slightly) more exploration.

A Bayesian model for increasing toxicities In the CRM literature, several parametric models that yield an increasing toxicity have been considered. In this paper, we choose a two-parameter logistic model that is among the most popular. Under this model, each dose k is assigned an *effective dose* u_k (that is usually not related to a true dose expressed in a mass or volume unit) and the toxicity probability of dose k is given by

$$p_k(\beta_0, \beta_1) = \psi(k, \beta_0, \beta_1), \quad \text{where} \quad \psi(k, \beta_0, \beta_1) = \frac{1}{1 + e^{-\beta_0 - \beta_1 u_k}}.$$

A typical choice of prior is

$$\beta_0 \sim \mathcal{N}(0, 100) \text{ and } \beta_1 \sim \text{Exp}(1).$$

It is worth noting that this model also heavily relies on the distinct effective dose levels $u_1, \ldots, u_K$ that are usually chosen depending on some *prior toxicities* set by physicians, $p_1^0 \leq p_2^0 \leq \cdots \leq p_K^0$. Letting $\overline{\beta_0}, \overline{\beta_1}$ be the prior mean of each parameter, the effective doses are calibrated such that for all k, $\psi(k, \overline{\beta_0}, \overline{\beta_1}) = p_k^0$. If

there is no medical prior knowledge about the toxicity probabilities, some heuristics for choosing them in a robust way have been developed (see Chapter 9 of Cheung (2011)).

Under this model, given some observations from the different doses one can compute the posterior distribution over the parameters β_0 and β_1; that is, the conditional distribution of these parameters given the observations. Although there is no closed form for these posterior distributions, they can be easily sampled from using Hamiltonian Monte-Carlo Markov Chain algorithms (HMC) as the log-likelihood under these models is differentiable. In practice, we use the Stan implementation of these Monte-Carlo sampler (Stan Development Team, 2015), and use (many) samples to approximate integrals under the posterior when needed.

4.1.1 Thompson Sampling

Thompson Sampling selects a dose at random according to its posterior probability of being the MTD. Under the two-parameter Bayesian logistic model presented above, letting π_t denote the posterior distribution on (β_0, β_1) after the first t observations, the posterior probability that dose k is the MTD is

$$\begin{aligned} q_k(t) &:= \mathbb{P}\left(k = \operatorname*{argmin}_{\ell} |\theta - p_\ell(\beta_0, \beta_1)| \middle| \mathcal{F}_t \right) \\ &= \int_{\mathbb{R}} \mathbb{1}\left(k = \operatorname*{argmin}_{\ell} |\theta - p_\ell(\beta_0, \beta_1)| \right) d\pi_t(\beta_0, \beta_1). \end{aligned}$$

A first possible implementation of Thompson Sampling that we use in our experiments consists of computing approximations $\hat{q}_k(t)$ of the probabilities $q_k(t)$ (using posterior samples) and selecting at round $t+1$ a dose $D_{t+1} \sim \hat{\boldsymbol{q}}(t)$, i.e. such that $\mathbb{P}\left(D_{t+1} = k | \mathcal{F}_t\right) = \hat{q}_k(t)$. A second implementation of Thompson Sampling (that may be computationally easier) consists of drawing one sample from the posterior distribution of (β_0, β_1), and selecting the MTD in the sampled model:

$$\begin{aligned} \left(\tilde{\beta}_0(t), \tilde{\beta}_1(t)\right) &\sim \pi_t, \\ D_{t+1}^{\mathrm{TS}} &\in \operatorname*{argmin}_{k \in \{1,\dots,K\}} \left| \theta - p_k\left(\tilde{\beta}_0(t), \tilde{\beta}_1(t)\right) \right|. \end{aligned} \tag{2}$$

It is easy to see that this algorithm coincides with Thompson Sampling in that $\mathbb{P}\left(D_{t+1}^{\mathrm{TS}} = k | \mathcal{F}_t\right) = q_k(t)$. We will present below a variant of Thompson Sampling based on the first implementation ($\mathrm{TS\_A}$) and a variant based on the second implementation ($\mathrm{TS}(\varepsilon)$).

Recommendation rule Due to the randomization, Thompson Sampling performs more exploration than the "greedy" CRM (O'Quigley et al., 1990) method, which selects at time t the MTD under the model parameterized by $(\hat{\beta}_0, \hat{\beta}_1)$, the posterior means of the two parameters, given by

$$\hat{\beta}_0(t) = \int_{\mathbb{R}} \beta_0 d\pi_t(\beta_0, \beta_1) \quad \text{and} \quad \hat{\beta}_1(t) = \int_{\mathbb{R}} \beta_1 d\pi_t(\beta_0, \beta_1). \tag{3}$$

More precisely, the sampling rule of the CRM is

$$D_{t+1}^{\mathrm{CRM}} \in \operatorname*{argmin}_{k \in \{1,\dots,K\}} \left| \theta - p_k(\hat{\beta}_0(t), \hat{\beta}_1(t)) \right|.$$

The recommendation rule for CRM after t patients is identical to the next dose that would be sampled under this design, that is $\hat{k}_t^{\text{CRM}} = D_{t+1}^{\text{CRM}}$. For Thompson Sampling, due to the more exploratory nature of the algorithm, we do not want to recommend $\hat{k}_t^{\text{TS}} = D_{t+1}^{\text{TS}}$. Instead, we propose the use of recommendation rule $\hat{k}_t^{\text{TS}} = \underset{k\in\{1,\dots,K\}}{\text{argmin}} \; |\theta - p_k(\hat{\beta}_0(t), \hat{\beta}_1(t))|$, which coincides with that of the CRM.

4.1.2 Two variants of Thompson Sampling

The randomized aspect of Thompson Sampling makes it likely to sample from large or small doses, without respecting some ethical constraints of phase I clinical trials. Indeed, patients should not be exposed to too-high dose levels; overdosing should be controlled. Hence, we also propose two "regularized" versions of TS. The first depends on a parameter $\varepsilon > 0$ set by the user that ensures that the expected toxicity of the recommended dose remains within ε of the toxicity of the empirical MTD. The second restricts the doses to be tested to a set of *admissible doses*. These algorithms are formally defined below, and their performance is evaluated in Section 5.

TS(ε) We first compute the posterior means $\hat{\beta}_0(t), \hat{\beta}_1(t)$ from (3) and the toxicity of the dose closest to θ under the model parameterized by $(\hat{\beta}_0(t), \hat{\beta}_1(t))$ (i.e., the toxicity of the dose selected by the CRM):

$$\hat{p}(t) = p_{\hat{k}_t}(\hat{\beta}_0(t), \hat{\beta}_1(t)), \quad \text{with} \quad \hat{k}_t = \underset{k\in\{1,\dots,K\}}{\text{argmin}} \left|\theta - p_k(\hat{\beta}_0(t), \hat{\beta}_1(t))\right|$$

Next we sample $\tilde{\beta}_0(t), \tilde{\beta}_1(t)$ from the posterior distribution π_t and select a candidate dose level D_{t+1} using (2). If the predicted toxicity level $p_{D_{t+1}}(\hat{\beta}_0(t), \hat{\beta}_1(t))$ is not in the interval $(\hat{p}(t) - \varepsilon, \hat{p}(t) + \varepsilon)$, then we reject our values of $\tilde{\beta}_0(t), \tilde{\beta}_1(t)$, draw a new sample from π_t and repeat the process. In order to guarantee that the algorithm terminates, we only reject up to 50 samples, after which we use the sample that gives the dose with minimum toxicity among all 50 samples. We choose 50 to limit the computational complexity of the algorithm, but it can also be replaced by a larger value if more computational power is available.

TS(ε) can be seen as a smooth interpolation between the CRM (which correspond to $\varepsilon = 0$) and vanilla Thompson Sampling (which corresponds to $\varepsilon = 1$). Regarding the tuning of the parameter ε, large values do not reduce much the amount of exploration while too small values lead to a behavior which is indistinguishable from that of the CRM. We did (large scale) experiments with $\varepsilon \in \{0.02, 0.05, 0.1\}$ and we found that the three values lead to comparable performance across the different scenarios we tried. To ease the presentation, we report results for TS(0.05) only in Section 5.

TS_A The TS_A algorithm limits exploration by enforcing the selected dose to be in some admissible set $\mathcal{A}_t$, by sampling from the modified distribution

$$\mathbb{P}\left(D_{t+1} = k | \mathcal{F}_t\right) = \frac{\hat{q}_k(t)\mathbb{1}_{(k\in\mathcal{A}_t)}}{\sum_{\ell\in\mathcal{A}_t}\hat{q}_\ell(t)},$$

instead of sampling directly from $\hat{\boldsymbol{q}}(t)$ as vanilla Thompson Sampling does. The admissible set $\mathcal{A}_t$ is defined as set of doses that meet the following two criteria:

1. dose k has either already been tested, or is the next-smallest dose which has not yet been tested

2. the posterior probability that the toxicity of dose k exceeds the toxicity of the dose closest to θ is smaller than some threshold:

$$\mathbb{P}\left(\psi(k,\beta_0,\beta_1) > \psi(k',\beta_0,\beta_1),\ \text{where } k' = \operatorname*{argmin}_{k'\in\{1,\dots,K\}} \left|\theta - \psi(k',\beta_0,\beta_1)\right| \middle| \mathcal{F}_t\right) \leq c_1.$$

$\mathcal{A}_t$ is inspired by the admissible set of Riviere et al. (2017) described in detail in the next section.

In our experiments, we tried different values of the parameter c_1 and we found that the performance of TS_A is better with values of c_1 that are not too small. In Section 5, we report experiments with $c_1 = 0.8$, but the performance was comparable for the choices $c_1 = 0.6$ or 0.9.

4.2 Thompson Sampling for Efficacy Plateau Models: A Phase I/II Design

In some particular trials, it has been established that efficacy is not always increasing with the dose. Motivated by some concrete examples discussed in their paper, Riviere et al. (2017) consider a model in which the dose effectiveness can plateau after some unknown level, while toxicity still increases with dose level. In these models, MTD identification is no longer relevant and the objective is rather to identify the smallest dose with maximal efficacy and with toxicity no more than θ. More formally, introducing eff_k the efficacy probability of dose k, the Minimal Effective Dose (MED) is

$$k^* = \min\left\{k : \text{eff}_k = \max_{\ell: p_\ell \leq \theta} \text{eff}_\ell\right\}$$

In a dose-finding study involving efficacy, at each time step t a dose D_t is allocated to the t-th patient, and the toxicity X_t is observed, as well as the efficacy Y_t. With this two-dimensional observation, assigning a value (or reward) to each sampled arm is even less natural than before. However as one can still define a notion of optimal dose (the MED instead of the MTD), Thompson Sampling can still be applied in this setting. As we shall see, it bears some similarities to the state-of-the-art method developed by Riviere et al. (2017).

A Bayesian model for toxicity and efficacy Thompson Sampling requires a Bayesian model for both the dose/toxicity and the dose/efficacy relationship that enforces an increasing toxicity and a increasing then plateau efficacy. We use the model proposed by Riviere et al. (2017), that we now describe.

Under this model, toxicity and efficacy are assumed to be independent. The (increasing) toxicity follows the two-dimensional Bayesian logistic model with effective doses u_k:

$$\begin{aligned} p_k = p_k(\beta_0,\beta_1) &= \psi(k,\beta_0,\beta_1) \\ \text{and}\quad \beta_0 \sim \mathcal{N}(0,100), &\quad \beta_1 \sim \text{Exp}(1). \end{aligned}$$

Efficacy also follows a logistic model, with an additional parameter τ that indicates the beginning of the plateau of efficacy. The efficacy probability of dose level k is

$$\text{eff}_k = \text{eff}_k(\gamma_0,\gamma_1,\tau) = \phi(k,\gamma_0,\gamma_1,\tau),\ \text{ where }\ \phi(k,\gamma_0,\gamma_1,\tau) := \frac{1}{1+e^{-[\gamma_0+\gamma_1(v_k\mathbb{1}(k<\tau)+v_\tau\mathbb{1}(k\geq\tau))]}},$$

with v_k the *effective efficacy* of dose k. Given $(t_1,\dots,t_K)$ such that $\sum_{i=1}^K t_i = 1$, a probability distribution on $\{1,\dots,K\}$, the three parameters (γ_0,γ_1,τ) are independent and drawn from the following prior distribution:

$$\gamma_0 \sim \mathcal{N}(0,100),\quad \gamma_1 \sim \text{Exp}(1),\quad \tau \sim (t_1,\dots,t_K).$$

The prior on τ may be provided by a physician or set to $(1/K, \dots, 1/K)$ in case one has no prior information. Just like the effective doses u_k (that we may now call effective toxicities), the effective efficacies v_k are calculated using prior efficacies $\text{eff}_1^0 \leq \cdots \leq \text{eff}_K^0$:

$$v_k = \left(\log \left(\frac{\text{eff}_k^0}{1 - \text{eff}_k^0} \right) - \overline{\gamma}_0 \right) \Big/ \overline{\gamma}_1,$$

where $\overline{\gamma}_0 = 0$ and $\overline{\gamma}_1 = 1$ are the prior means of the parameters γ_0 and γ_1.

Posterior sampling Let $\mathcal{D}_t^{\text{eff}} = \{(D_1, Y_1), \dots, (D_t, Y_t)\}$ be the efficacy data gathered in the first t rounds. Generating samples from the posterior distribution of $(\gamma_0, \gamma_1, \tau)$ given $\mathcal{D}_t^{\text{eff}}$ is a bit more involved than generating posterior samples from (β_0, β_1). Indeed, it cannot be handled directly with HMC given that (γ_0, γ_1) are continuous and τ is discrete. Thus, we proceed in the following way: we first draw samples from $p(\gamma_0, \gamma_1 | \mathcal{D}_t^{\text{eff}})$, which can be performed with HMC (and requires marginalizing out the discrete parameter τ, following the example of change point models given in the Stan manual (Stan Development Team, 2015)). Then we sample τ conditionally to $\gamma_0, \gamma_1, \mathcal{D}_t^{\text{eff}}$.

4.2.1 Thompson Sampling

Recall that the principle of Thompson Sampling is to randomly select doses according to their posterior probability of being optimal. This idea can also be applied in this more complex model, using the corresponding definition of optimality. Given a vector $\boldsymbol{\psi} = (\psi_1, \dots, \psi_K)$ of increasing toxicity probabilities and a vector $\boldsymbol{\phi} = (\phi_1, \dots, \phi_K)$ of increasing then plateau efficacy probabilities, the optimal dose is

$$\text{MED}(\boldsymbol{\psi}, \boldsymbol{\phi}) := \min \left\{ k : \phi_k = \max_{\ell : \psi_\ell \leq \theta} \phi_\ell \right\}.$$

The posterior probability of dose k to be optimal in that case is

$$q_k(t) := \mathbb{P}\left(k = \text{MED}\left(\psi(\cdot, \beta_0, \beta_1), \phi(\cdot, \gamma_0, \gamma_1, \tau) \right) \middle| \mathcal{F}_t \right)$$

and in our experiments, we implement Thompson Sampling by computing approximations $\hat{q}_k(t)$ from the quantities $q_k(t)$ (based on posterior samples) and then selecting a dose $D_{t+1} \sim \hat{\boldsymbol{q}}(t)$ where $\hat{\boldsymbol{q}}(t) = (\hat{q}_1(t), \dots, \hat{q}_K(t))$. Just like in the previous model, an alternative implementation of Thompson Sampling would sample parameters from their posterior distributions and select the optimal dose in this sampled model. Letting

$$\tilde{\beta}_0(t), \tilde{\beta}_1(t) \quad \text{and} \quad \tilde{\gamma}_0(t), \tilde{\gamma}_1(t), \tilde{\tau}(t),$$

be samples from the posterior distributions after t observations of the toxicity and efficacy parameters respectively, one can compute $\tilde{\psi}_k(t) = \psi(k, \tilde{\beta}_0(t), \tilde{\beta}_1(t))$ and $\tilde{\phi}_k(t) = \phi(k, \tilde{\gamma}_0(t), \tilde{\gamma}_1(t), \tilde{\tau}(t))$ for every dose k. Given the toxicity and efficacy vectors

$$\tilde{\boldsymbol{\psi}}(t) = \left(\tilde{\psi}_1(t), \dots, \tilde{\psi}_K(t) \right)$$
$$\text{and } \tilde{\boldsymbol{\phi}}(t) = \left(\tilde{\phi}_1(t), \dots, \tilde{\phi}_K(t) \right),$$

this implementation of Thompson Sampling selects at round $t+1$ $D_{t+1}^{\text{TS}} = \text{MED}\left(\tilde{\boldsymbol{\psi}}(t), \tilde{\boldsymbol{\phi}}(t) \right)$.

Recommendation rule Here also we expect Thompson Sampling to be too exploratory for dose recommendation. Hence, we base our recommendation on estimated values. Given the posterior means $\hat{\beta}_0(t), \hat{\beta}_1(t), \hat{\gamma}_0(t), \hat{\gamma}_1(t)$ (estimated from posterior samples) and $\hat{\tau}(t)$ the mode of the posterior distribution of the breakpoint (see the next section for its computation), we compute $\hat{\psi}_k(t) = \psi(k, \hat{\beta}_0(t), \hat{\beta}_1(t))$ and $\hat{\phi}_k(t) = \phi(k, \hat{\gamma}_0(t), \hat{\gamma}_1(t), \hat{\tau}(t))$ and recommend $\hat{k}_t = \text{MED}\left(\hat{\boldsymbol{\psi}}(t), \hat{\boldsymbol{\phi}}(t)\right)$.

4.2.2 A Variant of Thompson Sampling using Adaptive Randomization

Interestingly, the need for randomization in the context of plateau efficacy has already been observed by Riviere et al. (2017). More precisely, as we explain below, the algorithm MTA-RA described in that work can be viewed as an hybrid approach between Thompson Sampling and a CRM approach.

Additionally to the use of *adaptive randomization*, the MTA-RA algorithm also introduces a notion of *admissible set*. The set of admissible doses after t patients, denoted by $\mathcal{A}_t$, is the set of dose levels k meeting all of the following criteria:

1. dose k has either already been tested, or is the next-smallest dose which has not yet been tested
2. the posterior probability that the toxicity of dose k exceeds θ is smaller than some threshold:
$$\mathbb{P}\left(\psi(k, \beta_0, \beta_1) > \theta | \mathcal{F}_t\right) \leq c_1 \tag{4}$$
3. if the dose has been tested more than 3 times, the posterior probability that the efficacy is larger than ξ is larger than some threshold:
$$\mathbb{P}\left(\phi(k, \gamma_0, \gamma_1, \tau) > \xi | \mathcal{F}_t\right) \geq c_2 \tag{5}$$

Practical computation of the admissible set can be performed using posterior samples from (β_0, β_1) to check the criterion (4) and posterior samples from $(\gamma_0, \gamma_1, \tau)$ to check the criterion (5).

The MTA-RA algorithm works in two steps. The first step exploits the *posterior distribution of the breakpoint*, $t_k(t) := \mathbb{P}\left(\tau = k | \mathcal{D}_t^{\text{eff}}\right)$, and uses randomization to pick a value $\hat{\tau}(t)$ close to the mode of this distribution. More precisely, given $(\hat{t}_k(t))_{k=1,\dots,K}$ an estimate of the posterior distribution of τ, let

$$\mathcal{R}_t := \left\{ k : \left| \max_{1 \leq \ell \leq K} (\hat{t}_\ell(t)) - \hat{t}_k(t) \right| \leq s_1, 1 \leq k \leq K \right\}$$

be a set of candidate values for the position of the breakpoint. Then under MTA-RA,

$$\mathbb{P}\left(\hat{\tau}(t) = k | \mathcal{F}_t\right) = \frac{\hat{t}_k(t) \mathbb{1}_{(k \in \mathcal{R}_t)}}{\sum_{\ell \in \mathcal{R}_t} \hat{t}_\ell(t)}.$$

The threshold s_1 is often adapted such that it is larger in the beginning of the trial when we have high uncertainty about the estimates, but it grows smaller as the trial continues. The second step of MTA-RA doesn't employ randomization. Based on posterior samples from (γ_0, γ_1) conditionally to τ being equal to $\hat{\tau}(t)$, efficacy estimates $\hat{\phi}_k$ are produced (taking the mean of the values of $\phi(k, \tilde{\gamma_0}, \tilde{\gamma_1}, \hat{\tau}(t))$ for many samples $\tilde{\gamma_0}, \tilde{\gamma_1}$) and finally the selected dose is

$$D_{t+1}^{\text{MTA-RA}} = \inf \left\{ k \in \mathcal{A}_t : \hat{\phi}_k = \max_{j \in \mathcal{A}_t} \hat{\phi}_j \right\}.$$

If $\hat{\tau}(t)$ were replaced by a point estimate (e.g. the mode of the breakpoint posterior distribution $\hat{\boldsymbol{t}}(t)$), MTA-RA would be close to a CRM approach that computes estimates of all the parameters and acts greedily with respect to those estimated parameters (with the additional constraint that the chosen dose has to remain in the admissible set). However, the first step of MTA-RA bears similarities with the first step of a Thompson Sampling implementation that would sample a parameter τ from the $\hat{\boldsymbol{t}}(t)$ (and later sample the other parameters conditionally to that value and act greedily in the sampled model). The difference is the use of *adaptive* randomization, in which the sample is not exactly drawn from $\hat{\boldsymbol{t}}(t)$, but is constrained to fall in some set (here $\mathcal{R}_t$) that depends on previous observations.

The TS_A algorithm We believe that using adaptive randomization is a good idea to control the amount of exploration performed by Thompson Sampling, which leads us to propose the TS_A algorithm, that incorporates the constraint to select a dose that belongs to the admissible set $\mathcal{A}_t$. More formally, TS_A selects a dose at random according to

$$\mathbb{P}\left(D_{t+1} = k | \mathcal{F}_t\right) = \frac{\hat{q}_k(t)\mathbb{1}_{(k\in\mathcal{A}_t)}}{\sum_{\ell\in\mathcal{A}_t}\hat{q}_\ell(t)},$$

where we recall that $\hat{q}_k(t)$ is an estimate of the posterior probability that dose k is optimal. Compared to the variant of TS_A for increasing toxicities that is proposed in Section 4.1, the difference here is the appropriate definition of the admissible set, that involves both toxicity and efficacy probabilities.

Practical remark Approximations $\hat{t}_k(t)$ of the breakpoint distribution can be computed using that

$$t_k(t) = t_k \int \frac{L(\mathcal{D}_t^{eff}|\gamma_0,\gamma_1,k)}{\sum_{s=1}^K t_s L(\mathcal{D}_t^{eff}|\gamma_0,\gamma_1,s)} p(\gamma_0,\gamma_1|\mathcal{D}_t^{eff}) d\gamma_0 d\gamma_1,$$

where $L(\mathcal{D}_t^{\text{eff}}|\gamma_0,\gamma_1,s)$ is the likelihood of the efficacy observations when the efficacy model parameters are (γ_0,γ_1,s) and $p(\gamma_0,\gamma_1|\mathcal{D}_t^{\text{eff}})$ is the density of the distribution of (γ_0,γ_1) given the observations. $\hat{t}_k(t)$ can be thus be obtained by Monte-Carlo estimation based on samples from $p(\gamma_0,\gamma_1|\mathcal{D}_t^{\text{eff}})$.

5 Experimental Evaluation

We now present an empirical evaluation of the variants of Thompson Sampling introduced in the paper first in the context of increasing efficacy and then with the presence of a plateau of efficacy. In both groups of experiments, we adjusted our designs to some common practices in dose-finding trials. We used a start-up phase for all designs (starting from the smallest dose and escalating until the first toxicity is observed) and we also used cohorts of patients of size 3. This means that the same dose is allocated to 3 patients at a time and the model is updated after seeing the outcome for these 3 patients.

5.1 Phase I: MTD Identification

In this set of experiments, we evaluate the performance of the three algorithms introduced in Section 4.1, TS, TS(ε) and TS_A, and compare them to the 3+3 and CRM baselines. We report experiments with the value $\varepsilon = 0.05$ for TS(ε) and $c_1 = 0.8$ for TS_A. We refer the reader to Section 4.1.2 for discussions on the choice of these parameters. We also include Independent TS as proposed in Section 2, which is agnostic to the increasing structure.

In Tables 1 to 3 we provide results for nine different scenarios in which there are $K = 6$ doses with a target toxicity $\theta = 0.30$, budget $n = 36$ and prior toxicities

$$\boldsymbol{p^0} = [0.06 \;\; 0.12 \;\; 0.20 \;\; 0.30 \;\; 0.40 \;\; 0.50].$$

We choose the same prior toxicities for all scenario, that are sometimes close to actual toxicities (e.g. in Scenario 2) and sometimes quite far, in order to showcase the robustness of Bayesian algorithms.

For each scenario and algorithm, we report in the first column of these tables the percentage of allocation to each dose, that is, an estimate of $\mathbb{P}(\hat{k}_n = k)$ for each dose k, based on $N = 2000$ repetitions. In the second column, we report an estimate of the percentage of allocation to each dose during the trial, computed for each dose k as the average value of $100 * N_k(n)/n$ over $N = 2000$ repetitions. We add in parenthesis the empirical standard deviation of these allocation percentages, as allocations under bandit algorithms are known to have a large variance. For the 3+3 design, only the recommendation percentages are displayed, as the percentage of allocations would be computed based on a number of patients smaller than 36 (as a 3+3 based trial involves some random stopping). This design is also the only one that would stop and recommend none of the doses if they are all judged too toxic: we add this fraction of no recommendation between brackets in the tables.

For each scenario, corresponding to different increasing toxicity probabilities, the MTD is underlined and we mark in bold the fraction of recommendation or allocation of the MTD that are superior to what is achieved by the CRM. We now comment on the performance of the algorithms on those scenarios.

Dose recommendation TS outperforms CRM 3 out of 9 times, TS(ε) does so 5 out of 9 times, and TS_A does so 5 out of 9 times. As expected, Independent TS, which does not leverage the increasing structure, does not have a remarkable performance. This algorithm would need a larger budget to have a good empirical performance. With $n = 36$ in most cases this strategy is not doing much better than selecting the doses uniformly at random. One can also observe that the 3+3 design (that may however require less than 36 patients in the trial) performs very badly in terms of dose recommendation.

Dose allocation While TS_A and TS(ε) do not always have higher allocation percentage at the optimal (underlined) dose compared to CRM, a scan of the dose allocation results in Tables 1 to 3 shows that the addition of the admissible set $\mathcal{A}$ and ε regularity to the Thompson Sampling method consistently reduces the allocation percentage of higher toxicity doses. TS_A performs best in this regard (it is more cautious with allocating higher doses) across all algorithms (e.g. it consistently has superior performance compared to CRM), while TS(ε) has performance better than or comparable to CRM. We believe this result is of interest in trials where toxicity is an ethical concern.

5.2 Phase I/II: MED Identification when Efficacy Plateaus

In this set of experiments, we evaluate the performance of the two algorithms introduced in Section 4.2, TS and TS_A, and compare them to the MTA-RA algorithm. We use the experimental setup of Riviere et al. (2017): several scenarios with $K = 6$ doses, budget $n = 60$, $\theta = 0.35$, toxicity and efficacy priors

$$\boldsymbol{p^0} = [0.02, 0.06, 0.12, 0.20, 0.30, 0.40] \quad \text{and} \quad \mathbf{eff^0} = [0.12, 0.20, 0.30, 0.40, 0.50, 0.59].$$

Furthermore, we use the same parameters for the admissible set and the implementation of MTA-RA as those chosen by Riviere et al. (2017): $\xi = 0.2$, $c_1 = 0.9$, $c_2 = 0.4$, and $s_1 = .2\left(1 - \frac{I}{n}\right)$, where I is the number of samples used so far. These parameters are defined above in the main text.

In Tables 4 to 6 we provide results for several scenarios with increasing toxicities and efficacy, with efficacy which (quasi) plateaus. We report the percentage of allocation to each dose, the percentage of recommendation of each dose when $n = 60$, and the percentage of time the trials stopped early (E-Stop), estimated over $N = 2000$ repetitions. As before, we also report standard deviations for the percentage of allocations to each dose.

Optimal doses are underlined by a plain line while a dashed line identifies doses whose toxicity is larger than θ. We mark in bold cases where our algorithms makes the optimal decision (in terms of the percentage of recommendations) more often than the MTA-RA baseline.

Dose recommendation Recall that the modeling assumption here is that efficacy increases monotonically in toxicity up to a point and then it plateaus. We present experimental results on several scenarios, some of which are borrowed from Riviere et al. (2017), on which this plateau assumption is not always exactly met. In most of these scenarios, TS_A outperforms the MTA-RA algorithm.

In scenarios 1 through 4 and in scenarios 12 and 13, there is a plateau of efficacy starting at a reasonable toxicity: in this case the optimal dose corresponds to the plateau breakpoint. Our algorithms make the optimal decision compared to MTA-RA consistently: TS 4 out of 6 times and TS_A 5 out of 6 times. In scenarios 5 and 6 the plateau of efficacy starts when the toxicity is already too high, hence the optimal dose is before than the plateau. In scenario 5, TS_A and TS both outperform MTA-RA, while on scenario 6 MTA-RA has a slight advantage over TS.

In scenario 7 and 8 there is no true plateau of efficacy, however in both cases there exists a "breakpoint" (underlined) after which the efficacy is increasing very slowly while the toxicity is increasing significantly. This breakpoint can thus be argued to be a good trade-off between efficacy and toxicity and should be investigated in further phases. In these two scenarios TS_A identifies this pseudo-optimal dose more often than MTA-RA, while TS has a slightly worse performance.

Lastly, we study the case when there is no clear optimal or near-optimal dose, i.e. scenarios 9-11. In scenario 9 wherein most doses, including the entire quasi-plateau, are too toxic, we would like to stop early or at most recommend dose 1 (the only dose meeting the toxicity constraint but whose efficacy is not very high). Under this interpretation, TS and TS_A outperform MTA-RA. Note that our algorithms most often either stop early or recommend dose 1, while in comparison MTA-RA recommends the toxic dose 2 a large fraction of the time (33.1 %). In scenarios 10 and 11 in which all doses are either too toxic or ineffective a good algorithm would stop early with no recommendation. TS_A makes this optimal decision more often than MTA-RA in both scenarios and TS in one of the two scenarios.

Dose allocation While TS and TS_A have lower allocation percentage at the optimal (underlined) dose compared to MTA-RA, the addition of the admissible set $\mathcal{A}$ to the Thompson Sampling method consistently reduces the percentage of dose allocation at doses that are too toxic. Furthermore, TS_A is more cautious in allocating higher doses compared to MTA-RA. Our experiments notably reveal that the fraction of allocation to doses whose toxicity is larger than θ (that are underlined with a dashed line) is always smaller for TS_A than for MTA-RA. Hence, not only is TS_A very good in terms of recommending the right dose, it also manages to avoid too-toxic doses more consistently.

6 Revisiting the Treatment versus Experimentation Trade-off

Ideally, a good design for MTD identification should be supported by a control of both the error probability $e_n = \mathbb{P}(\hat{k}_n \neq k^*)$ and the number of sub-optimal selections $\mathbb{E}[N_k(n)]$ for $k \neq k^*$. These two quantities

Table 1: Results for MTD identification

Algorithm	Recommended						Allocated					
	1	2	3	4	5	6	1	2	3	4	5	6
Sc. 1: Tox prob	0.30	0.45	0.55	0.60	0.75	0.80	0.30	0.45	0.55	0.60	0.75	0.80
3 + 3 [30.4]	35.2	21.6	7.7	4.0	1.0	0.1	-	-	-	-	-	-
CRM	77.2	20.8	1.9	0.1	0.0	0.0	70.1	21.7	6.2	1.5	0.3	0.3
							(32.1)	(24.1)	(12.1)	(5.4)	(1.9)	(1.7)
TS	**78.9**	18.9	2.2	0.0	0.0	0.0	67.0	18.8	6.3	2.3	1.0	4.6
							(24.4)	(16.2)	(9.5)	(5.2)	(3.1)	(5.6)
TS(ϵ)	**78.6**	19.5	1.8	0.1	0.0	0.0	**73.0**	20.7	5.2	0.7	0.1	0.3
							(31.0)	(24.3)	(10.8)	(3.6)	(1.2)	(1.6)
TS_A	**79.8**	18.2	1.7	0.2	0.1	0.0	**76.3**	19.7	3.5	0.5	0.1	0.0
							(24.1)	(18.7)	(8.7)	(3.2)	(1.0)	(0.3)
Independent TS	37.6	27.3	16.9	13.2	3.2	1.9	23.4	21.0	17.4	15.9	11.7	10.6
							(12.9)	(11.5)	(10.2)	(9.5)	(6.2)	(5.5)
Sc. 2: Tox prob	0.05	0.12	0.15	0.30	0.45	0.50	0.05	0.12	0.15	0.30	0.45	0.50
3 + 3 [0.5]	4.9	6.2	23.0	28.6	18.8	17.9	-	-	-	-	-	-
CRM	0.2	1.2	17.1	53.9	21.7	5.9	10.3	10.7	20.6	29.9	15.9	12.7
							(6.4)	(11.1)	(19.6)	(21.0)	(16.3)	(16.6)
TS	0.0	1.2	17.8	47.2	25.9	8.0	13.6	14.6	18.2	20.3	12.5	20.7
							(8.8)	(10.6)	(13.0)	(13.8)	(11.2)	(15.2)
TS(ϵ)	0.2	1.5	14.9	51.5	24.2	7.6	10.4	11.1	22.3	**30.2**	13.0	13.0
							(6.8)	(11.3)	(19.6)	(21.2)	(15.3)	(16.0)
TS_A	0.0	1.8	14.9	44.3	24.9	14.1	15.3	19.5	25.7	23.9	10.2	5.5
							(11.6)	(14.6)	(15.5)	(17.0)	(12.3)	(11.1)
Independent TS	17.7	17.2	19.2	20.2	14.7	11.0	16.0	18.6	18.7	17.6	15.0	14.2
							(8.8)	(7.5)	(8.0)	(8.9)	(8.3)	(8.0)
Sc. 3: Tox prob	0.01	0.03	0.07	0.11	0.15	0.30	0.01	0.03	0.07	0.11	0.15	0.30
3 + 3 [0.0]	0.3	1.8	3.6	6.2	20.8	67.2	-	-	-	-	-	-
CRM	9.6	0.0	0.1	1.4	14.8	74.1	14.0	8.2	8.9	8.7	14.8	45.4
							(15.1)	(1.8)	(4.5)	(8.2)	(14.5)	(21.4)
TS	2.9	0.0	0.1	1.8	14.8	**80.2**	11.8	9.2	10.1	11.7	14.1	43.2
							(9.8)	(3.6)	(6.2)	(8.8)	(10.5)	(16.3)
TS(ϵ)	2.9	0.1	0.1	1.5	15.8	**79.8**	11.0	8.4	9.0	10.5	15.3	**45.8**
							(8.9)	(2.5)	(4.7)	(9.2)	(13.9)	(19.1)
TS_A	2.5	0.0	0.1	1.7	14.3	**81.5**	11.7	10.6	13.6	15.9	16.0	32.1
							(9.4)	(6.2)	(9.7)	(10.9)	(10.6)	(19.0)
Independent TS	18.8	10.0	14.4	19.4	18.6	19.0	15.3	16.3	16.8	17.6	17.6	16.4
							(8.2)	(5.4)	(6.3)	(7.1)	(7.5)	(8.2)

Table 2: Results for MTD identification (part 2/3)

Algorithm	Recommended						Allocated					
	1	2	3	4	5	6	1	2	3	4	5	6
Sc. 4: Tox prob	0.10	0.20	0.30	0.40	0.47	0.53	0.10	0.20	0.30	0.40	0.47	0.53
3 + 3 [4.7]	12.5	20.5	23.0	18.8	11.6	9.0	-	-	-	-	-	-
CRM	1.2	22.0	42.2	25.7	6.9	2.1	14.6 (13.7)	23.1 (23.7)	30.6 (24.0)	18.0 (19.4)	7.4 (12.3)	6.3 (12.2)
TS	1.2	19.6	40.1	28.2	8.1	2.8	21.4 (15.1)	21.6 (15.4)	20.8 (15.0)	13.7 (12.3)	6.8 (8.7)	15.7 (13.6)
TS(ϵ)	2.1	19.9	**44.1**	24.9	7.0	1.8	15.5 (15.8)	25.3 (24.5)	31.8 (24.5)	16.2 (19.2)	5.1 (9.8)	6.1 (10.8)
TS_A	1.4	20.6	**42.3**	22.2	9.0	4.5	25.1 (19.3)	31.0 (18.4)	27.4 (18.8)	11.9 (14.9)	3.2 (7.9)	1.3 (5.6)
Independent TS	17.8	22.2	22.6	15.9	12.6	9.0	16.6 (9.3)	19.4 (8.7)	18.7 (9.3)	16.5 (8.8)	15.3 (8.6)	13.5 (7.7)
Sc. 5: Tox prob	0.10	0.25	0.40	0.50	0.65	0.75	0.10	0.25	0.40	0.50	0.65	0.75
3 + 3 [3.1]	20.6	30.8	24.2	15.3	5.1	0.8	-	-	-	-	-	-
CRM	4.8	49.7	39.0	6.5	0.1	0.0	17.8 (18.2)	38.3 (27.4)	30.9 (23.9)	9.0 (14.8)	2.4 (5.5)	1.7 (4.0)
TS	4.3	**50.7**	39.4	5.4	0.1	0.1	26.3 (17.6)	31.2 (17.5)	22.3 (16.0)	8.8 (11.4)	3.2 (5.4)	8.2 (7.2)
TS(ϵ)	4.8	**52.2**	36.5	6.2	0.2	0.0	18.8 (19.3)	**41.2** (27.1)	29.7 (24.4)	7.3 (13.7)	1.4 (4.2)	1.6 (3.9)
TS_A	3.0	**50.8**	36.4	7.0	1.6	1.1	29.6 (20.0)	**40.1** (18.8)	23.4 (18.5)	6.1 (11.0)	0.8 (3.2)	0.1 (1.1)
Independent TS	24.3	32.6	21.4	14.6	5.4	1.6	19.4 (10.5)	22.6 (10.8)	19.1 (10.0)	16.0 (9.1)	12.5 (7.0)	10.4 (5.5)
Sc. 6: Tox prob	0.08	0.12	0.18	0.25	0.33	0.39	0.08	0.12	0.18	0.25	0.33	0.39
3 + 3 [2.1]	5.1	9.6	15.3	19.3	18.5	30.2	-	-	-	-	-	-
CRM	0.3	1.2	10.6	29.1	31.2	27.5	11.7 (8.4)	10.7 (11.4)	16.2 (17.1)	19.5 (19.3)	18.2 (18.0)	23.7 (23.9)
TS	0.3	1.4	10.6	27.0	29.9	30.9	14.9 (10.7)	13.2 (9.9)	15.3 (12.0)	15.1 (12.3)	12.2 (11.3)	29.2 (18.7)
TS(ϵ)	0.1	1.7	11.5	28.3	30.4	28.0	12.0 (8.9)	11.9 (12.5)	19.2 (18.6)	**20.3** (19.2)	13.5 (15.4)	23.0 (22.7)
TS_A	0.1	1.9	12.0	28.5	26.5	31.0	17.5 (14.4)	21.1 (15.0)	24.7 (15.9)	19.3 (15.9)	8.9 (11.5)	8.5 (15.0)
Independent TS	13.6	15.6	19.1	19.4	16.8	15.4	14.7 (8.4)	17.7 (7.4)	18.0 (8.0)	17.5 (8.5)	16.4 (8.5)	15.7 (8.4)

Table 3: Results for MTD identification (part 3/3)

Algorithm	Recommended						Allocated					
	1	2	3	4	5	6	1	2	3	4	5	6
Sc. 7: Tox prob	0.15	0.30	0.45	0.50	0.60	0.70	0.15	0.30	0.45	0.50	0.60	0.70
3 + 3 [7.7]	24.7	32.8	18.0	10.2	4.9	1.8	-	-	-	-	-	-
CRM	16.9	59.4	20.4	3.0	0.2	0.2	27.7 (27.2)	40.8 (27.1)	22.4 (22.1)	6.0 (11.6)	1.8 (5.5)	1.4 (4.2)
TS	14.5	55.7	25.6	3.9	0.1	0.1	34.9 (21.8)	29.5 (17.0)	17.5 (14.8)	7.0 (9.8)	2.9 (5.4)	8.2 (8.0)
TS(ϵ)	15.0	58.0	23.2	3.5	0.2	0.1	28.8 (27.5)	**43.3** (27.0)	20.8 (21.9)	4.7 (11.0)	1.0 (4.0)	1.5 (4.0)
TS_A	13.7	**59.5**	21.5	3.7	0.9	0.8	41.7 (24.8)	39.3 (18.9)	15.5 (16.8)	3.1 (7.9)	0.4 (2.7)	0.1 (1.2)
Independent TS	25.4	33.1	16.8	13.7	7.6	3.4	19.2 (11.0)	22.5 (11.1)	17.5 (9.9)	16.3 (9.4)	13.3 (7.7)	11.2 (6.3)
Sc. 8: Tox prob	0.10	0.15	0.30	0.45	0.60	0.75	0.10	0.15	0.30	0.45	0.60	0.75
3 + 3 [3.1]	6.8	24.1	30.8	22.4	11.0	1.8	-	-	-	-	-	-
CRM	1.1	15.1	60.6	21.6	1.6	0.1	13.5 (12.4)	20.4 (20.8)	39.6 (24.5)	18.4 (20.2)	4.9 (9.3)	3.1 (5.5)
TS	0.9	21.0	58.5	18.5	1.0	0.1	20.4 (15.4)	23.8 (15.0)	27.4 (16.3)	13.8 (13.5)	4.9 (7.3)	9.8 (7.6)
TS(ϵ)	0.8	17.0	59.4	20.2	2.0	0.7	14.4 (14.0)	23.4 (21.8)	**39.9** (24.3)	16.3 (19.8)	2.6 (6.1)	3.4 (5.4)
TS_A	0.3	14.5	51.9	24.0	5.4	3.9	22.4 (17.5)	30.1 (17.6)	31.7 (18.3)	13.0 (14.7)	2.3 (6.0)	0.5 (2.3)
Independent TS	22.4	24.4	26.8	17.1	7.4	2.0	18.3 (9.9)	21.5 (9.1)	20.5 (9.8)	16.9 (9.4)	13.0 (7.4)	9.9 (5.3)
Sc. 9: Tox prob	0.01	0.05	0.08	0.15	0.30	0.45	0.01	0.05	0.08	0.15	0.30	0.45
3 + 3 [0.1]	0.8	2.1	8.0	23.8	30.0	35.2	-	-	-	-	-	-
CRM	1.9	0.1	0.4	16.1	54.1	27.4	9.8 (7.3)	8.5 (3.5)	10.0 (7.6)	17.0 (16.4)	28.9 (18.9)	25.8 (20.8)
TS	0.5	0.0	0.5	17.1	50.8	31.1	10.3 (6.0)	10.0 (4.8)	12.0 (7.9)	18.3 (12.1)	20.0 (12.7)	29.4 (16.0)
TS(ϵ)	0.7	0.1	0.4	15.2	**55.9**	27.9	9.3 (5.3)	8.4 (2.5)	10.6 (7.3)	20.2 (17.2)	26.3 (18.3)	25.2 (19.9)
TS_A	0.3	0.0	0.5	13.2	46.7	39.2	10.4 (5.8)	12.3 (8.4)	16.5 (11.5)	22.9 (14.2)	19.9 (13.3)	18.1 (17.1)
Independent TS	18.8	11.6	14.8	19.0	21.0	14.8	15.4 (8.4)	17.1 (6.2)	17.5 (6.8)	18.1 (7.6)	17.1 (8.4)	14.9 (7.9)

Table 4: Results for MED identification (part 1/3).

Algorithm	E-Stop	Recommended						Allocated					
		1	2	3	4	5	6	1	2	3	4	5	6
Sc. 1: Tox prob		0.01	0.05	0.15	0.2	0.45	0.6	0.01	0.05	0.15	0.2	0.45	0.6
Sc. 1: Eff prob		0.1	0.35	0.6	0.6	0.6	0.6	0.1	0.35	0.6	0.6	0.6	0.6
MTA-RA	0.4	0.4	7.0	54.9	29.1	7.4	0.8	7.1 (3.8)	14.2 (13.9)	37.9 (24.4)	24.9 (18.8)	12.9 (13.6)	2.5 (4.9)
TS	0.9	0.1	9.7	**57.6**	27.0	4.2	0.4	10.6 (5.7)	18.4 (11.0)	31.9 (14.4)	23.8 (13.2)	10.0 (8.0)	4.4 (4.5)
TS_A	0.9	0.3	9.6	**59.4**	26.1	3.5	0.2	10.7 (5.4)	20.7 (12.9)	35.7 (14.9)	23.9 (14.1)	7.3 (8.1)	0.9 (2.7)
Sc. 2: Tox prob		0.005	0.01	0.02	0.05	0.1	0.15	0.005	0.01	0.02	0.05	0.1	0.15
Sc. 2: Eff prob		0.001	0.1	0.3	0.5	0.8	0.8	0.001	0.1	0.3	0.5	0.8	0.8
MTA-RA	1.9	0.0	0.1	1.6	5.1	55.0	36.2	5.2 (1.7)	5.6 (3.1)	7.5 (8.5)	11.4 (13.6)	36.7 (25.8)	31.7 (26.9)
TS	0.8	0.0	0.0	0.5	4.7	**56.6**	37.5	5.9 (2.4)	6.6 (3.4)	9.3 (6.0)	16.9 (9.7)	32.5 (13.3)	28.1 (14.4)
TS_A	2.2	0.0	0.1	1.6	5.0	**55.9**	35.2	5.9 (2.3)	6.8 (3.8)	10.9 (8.7)	17.9 (10.8)	31.8 (14.3)	24.5 (15.5)
Sc. 3: Tox prob		0.01	0.05	0.1	0.25	0.5	0.7	0.01	0.05	0.1	0.25	0.5	0.7
Sc. 3: Eff prob		0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
MTA-RA	0.4	51.5	26.4	12.5	6.8	2.2	0.2	38.2 (25.2)	24.8 (17.9)	16.6 (13.9)	12.9 (12.3)	6.1 (8.4)	0.9 (2.7)
TS	0.1	**53.9**	24.8	12.2	7.8	1.1	0.1	24.1 (11.4)	22.7 (9.8)	23.8 (10.9)	19.0 (10.6)	7.2 (6.1)	3.1 (3.6)
TS_A	0.5	**53.8**	26.4	10.4	8.2	0.7	0.1	26.6 (13.3)	25.1 (11.4)	24.8 (11.4)	17.7 (11.9)	4.8 (6.6)	0.5 (2.0)
Sc. 4: Tox prob		0.01	0.02	0.05	0.1	0.2	0.3	0.01	0.02	0.05	0.1	0.2	0.3
Sc. 4: Eff prob		0.25	0.45	0.65	0.65	0.65	0.65	0.25	0.45	0.65	0.65	0.65	0.65
MTA-RA	0.1	1.8	13.2	49.0	21.7	8.5	5.7	9.5 (7.9)	17.7 (15.9)	31.6 (21.5)	20.6 (15.6)	13.9 (12.6)	6.6 (10.2)
TS	0.1	1.8	15.7	45.8	18.1	10.8	7.8	12.1 (6.8)	16.8 (8.9)	23.1 (11.0)	21.6 (10.2)	16.5 (9.3)	9.8 (7.6)
TS_A	0.2	2.4	15.0	**49.1**	20.2	9.8	3.2	13.2 (8.0)	19.3 (10.8)	25.5 (12.3)	21.9 (10.8)	14.1 (10.7)	5.8 (7.9)
Sc. 5: Tox prob		0.1	0.2	0.25	0.4	0.5	0.6	0.1	0.2	0.25	0.4	0.5	0.6
Sc. 5: Eff prob		0.3	0.4	0.5	0.7	0.7	0.7	0.3	0.4	0.5	0.7	0.7	0.7
MTA-RA	1.4	9.0	13.2	25.9	40.6	8.3	1.5	15.5 (16.7)	19.1 (17.0)	24.9 (17.7)	26.7 (19.5)	9.9 (11.0)	2.4 (5.0)
TS	5.8	8.3	24.4	**40.0**	18.9	2.4	0.3	20.8 (15.8)	27.3 (15.8)	24.4 (14.8)	13.0 (11.2)	5.5 (6.3)	3.3 (4.3)
TS_A	6.9	16.7	30.6	**30.6**	14.4	0.8	0.0	25.9 (19.2)	33.8 (18.6)	22.8 (16.1)	8.6 (11.7)	1.8 (4.8)	0.2 (1.3)

Table 5: Results for MED identification (part 2/3).

Algorithm	E-Stop	Recommended						Allocated					
		1	2	3	4	5	6	1	2	3	4	5	6
Sc. 6: Tox prob		0.1	0.3	0.35	0.4	0.5	0.6	0.1	0.3	0.35	0.4	0.5	0.6
Sc. 6: Eff prob		0.3	0.4	0.5	0.7	0.7	0.7	0.3	0.4	0.5	0.7	0.7	0.7
MTA-RA	4.2	11.2	24.3	24.6	28.9	5.4	1.3	17.9 (19.9)	24.2 (22.2)	23.7 (18.6)	20.7 (19.8)	7.7 (10.5)	1.7 (4.2)
TS	8.4	17.8	41.9	22.4	8.1	1.1	0.2	29.6 (21.3)	30.4 (17.0)	16.9 (13.4)	8.2 (9.5)	4.0 (5.7)	2.4 (3.8)
TS_A	9.4	28.5	43.6	14.2	4.0	0.2	0.0	34.5 (24.0)	37.2 (20.2)	14.3 (14.5)	3.9 (8.3)	0.6 (2.7)	0.1 (0.9)
Sc. 7: Tox prob		0.03	0.06	0.1	0.2	0.4	0.5	0.03	0.06	0.1	0.2	0.4	0.5
Sc. 7: Eff prob		0.3	0.5	0.52	0.54	0.55	0.55	0.3	0.5	0.52	0.54	0.55	0.55
MTA-RA	0.1	8.6	45.5	25.1	13.7	5.7	1.4	16.1 (14.6)	31.5 (21.8)	22.8 (17.0)	17.0 (14.0)	9.9 (10.9)	2.5 (6.1)
TS	0.7	10.3	43.7	22.1	16.3	5.7	1.2	17.5 (8.9)	22.7 (11.3)	23.2 (10.8)	20.6 (11.0)	10.3 (7.6)	4.9 (5.2)
TS_A	0.4	11.3	**47.9**	22.8	13.3	4.2	0.1	19.8 (11.2)	26.9 (13.3)	26.1 (11.6)	19.0 (12.6)	6.6 (8.0)	1.2 (3.5)
Sc. 8: Tox prob		0.02	0.07	0.13	0.17	0.25	0.3	0.02	0.07	0.13	0.17	0.25	0.3
Sc. 8: Eff prob		0.3	0.5	0.7	0.73	0.76	0.77	0.3	0.5	0.7	0.73	0.76	0.77
MTA-RA	0.1	1.1	10.2	39.0	24.4	16.8	8.4	9.3 (7.5)	15.8 (14.8)	28.8 (21.0)	22.6 (16.0)	15.7 (14.1)	7.8 (12.3)
TS	0.3	1.2	11.1	36.9	24.2	16.1	10.2	12.1 (7.2)	17.4 (9.9)	24.1 (12.1)	21.9 (10.8)	15.0 (10.3)	9.1 (8.2)
TS_A	0.3	1.8	13.2	**45.6**	24.1	11.4	3.7	14.2 (9.4)	22.2 (13.5)	28.6 (13.7)	21.0 (12.5)	10.3 (11.2)	3.4 (6.9)
Sc. 9: Tox prob		0.25	0.43	0.50	0.58	0.64	0.75	0.25	0.43	0.50	0.58	0.64	0.75
Sc. 9: Eff prob		0.3	0.4	0.5	0.6	0.61	0.63	0.3	0.4	0.5	0.6	0.61	0.63
MTA-RA	18.8	40.0	33.1	7.0	0.9	0.1	0.1	32.0 (30.3)	30.3 (24.5)	13.6 (15.6)	4.3 (7.4)	1.0 (3.2)	0.1 (0.7)
TS	**49.0**	37.3	12.4	1.1	0.1	0.0	0.0	29.0 (31.9)	13.7 (16.5)	4.5 (7.5)	1.9 (3.9)	1.1 (2.8)	0.8 (2.1)
TS_A	**50.5**	39.8	9.2	0.5	0.1	0.0	0.0	31.2 (35.1)	14.6 (18.6)	3.3 (7.0)	0.4 (1.9)	0.0 (0.3)	0.0 (0.2)
Sc. 10: Tox prob		0.05	0.1	0.25	0.55	0.7	0.9	0.05	0.1	0.25	0.55	0.7	0.9
Sc. 10: Eff prob		0.01	0.02	0.05	0.35	0.55	0.7	0.01	0.02	0.05	0.35	0.55	0.7
MTA-RA	91.8	0.5	0.5	2.3	4.8	0.1	0.0	0.6 (2.8)	0.7 (2.6)	1.3 (5.9)	4.4 (15.4)	1.1 (4.3)	0.2 (1.1)
TS	61.9	12.2	2.5	1.8	19.7	1.8	0.1	3.8 (6.8)	8.9 (15.2)	16.3 (23.0)	6.3 (10.6)	1.9 (4.0)	0.9 (2.3)
TS_A	**94.1**	0.2	0.1	1.2	4.3	0.1	0.0	0.5 (2.2)	0.6 (2.5)	1.4 (6.1)	2.9 (12.0)	0.5 (2.4)	0.0 (0.4)

Table 6: Results for MED identification (part 3/3).

Algorithm	E-Stop	Recommended						Allocated					
		1	2	3	4	5	6	1	2	3	4	5	6
Sc. 11: Tox prob		0.5	0.6	0.69	0.76	0.82	0.89	0.5	0.6	0.69	0.76	0.82	0.89
Sc. 11: Eff prob		0.4	0.55	0.65	0.65	0.65	0.65	0.4	0.55	0.65	0.65	0.65	0.65
MTA-RA	90.1	9.6	0.2	0.1	0.0	0.0	0.0	7.2 (22.7)	2.0 (7.9)	0.5 (2.3)	0.1 (0.9)	0.0 (0.2)	0.0 (0.0)
TS	**99.8**	0.2	0.0	0.0	0.0	0.0	0.0	0.1 (3.0)	0.1 (1.2)	0.0 (0.4)	0.0 (0.1)	0.0 (0.0)	0.0 (0.1)
TS_A	**99.5**	0.5	0.0	0.0	0.0	0.0	0.0	0.4 (5.6)	0.1 (1.6)	0.0 (0.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Sc. 12: Tox prob		0.01	0.02	0.05	0.1	0.25	0.5	0.01	0.02	0.05	0.1	0.25	0.5
Sc. 12: Eff prob		0.05	0.25	0.45	0.7	0.7	0.7	0.05	0.25	0.45	0.7	0.7	0.7
MTA-RA	1.0	0.1	1.2	8.9	52.8	29.4	6.4	5.8 (2.4)	7.6 (6.5)	14.6 (15.7)	35.9 (24.2)	24.9 (20.0)	10.2 (12.8)
TS	0.8	0.0	0.7	10.0	**57.0**	27.7	4.0	7.7 (4.2)	10.4 (6.5)	17.9 (10.1)	32.2 (13.6)	21.9 (11.9)	9.3 (7.0)
TS_A	1.7	0.0	1.4	10.0	**56.0**	26.8	4.2	7.5 (3.9)	11.3 (8.5)	19.5 (11.3)	32.1 (14.4)	21.6 (13.0)	6.4 (7.5)
Sc. 13: Tox prob		0.01	0.05	0.1	0.2	0.3	0.5	0.01	0.05	0.1	0.2	0.3	0.5
Sc. 13: Eff prob		0.05	0.1	0.2	0.35	0.55	0.55	0.05	0.1	0.2	0.35	0.55	0.55
MTA-RA	14.9	0.7	1.8	5.5	17.0	50.3	9.7	6.4 (6.5)	7.4 (7.3)	11.1 (12.6)	18.7 (18.7)	30.7 (23.8)	10.8 (14.0)
TS	8.6	0.5	1.8	6.7	37.6	39.0	5.6	9.1 (6.3)	11.5 (8.2)	17.5 (12.0)	26.3 (14.8)	18.6 (14.2)	8.4 (7.7)
TS_A	17.3	0.5	1.4	7.4	31.6	37.5	4.2	7.2 (4.5)	9.1 (6.8)	16.7 (13.7)	26.8 (17.1)	18.1 (15.3)	4.7 (7.3)

are respectively useful to check whether the design achieves a *good identification of the optimal dose* and whether *a large number of patients have been treated with the optimal dose*.

For classical bandits (in which k^* is the arm with largest mean instead of the MTD), those two performance measures are known to be antagonistic. Indeed, Bubeck et al. (2011) shows that the smaller the regret (a quantity that can be related to the number of sub-optimal selections), the larger the error probability. Such a trade-off may also exist for the MTD identification problem. However, the precise statement of such a result would be meaningful for large values of the number of patients n, which is of little interest for a real clinical trial as it can only involve a small number of patients. In practice, we showed that adaptations of Thompson Sampling, a bandit design aimed at maximizing rewards, achieve good performance in terms of both allocation and recommendation.

Still, another natural avenue of research is to investigate the adaptation of bandit designs aimed at minimizing the error probability. Minimizing the error probability for MTD can be viewed as a variant of the fixed-budget Best Arm Identification (BAI) problem introduced by Audibert et al. (2010); Bubeck et al. (2011). In contrast to the standard BAI problem that aims to identify the arm with largest mean (which would correspond here to the most toxic dose), the focus is on identifying the arm whose mean is closest to the threshold θ. A state-of-the art fixed-budget BAI algorithm is Sequential Halving (Karnin et al., 2013), and we propose in Algorithm 6 a natural adaptation to MTD identification.

Sequential Halving for MTD identification proceeds in phases. In each of the $\log_2(K)$ phases, all the remaining doses are allocated the same amount of times to patients and their empirical toxicity based on these allocations (that is, the average of the toxicity responses) is computed. At the end of each phase the empirical worst half of the doses is eliminated. For MTD identification, rather than the doses with the smallest empirical means (as the vanilla Sequential Halving algorithm would do), the doses whose empirical toxicity are the furthest away from the threshold θ are eliminated. Observe that by design of the algorithm, the total number of allocated doses is indeed smaller than the prescribed budget n.

Algorithm 1 Sequential Halving for MTD Identification

Input: budget n, target toxicity θ
Initialization: Set of dose levels $S_0 \leftarrow \{1, \dots, K\}$;
for $r \leftarrow 0$ to $\lceil \log_2(K) \rceil - 1$ **do**
  Allocate each dose $k \in S_r$ to $t_r = \left\lfloor \frac{n}{|S_r| \lceil \log_2(K) \rceil} \right\rfloor$ patients;
  Based on their response compute $\hat{p}_k^r$, the empirical toxicity of dose k based on these t_r samples;
  Compute S_{r+1} the set of $\lceil |S_r|/2 \rceil$ arms with smallest $\hat{d}_k^r := |\theta - \hat{p}_k^r|$
Output: the unique arm in $S_{\lceil \log_2(K) \rceil}$

Building on the analysis of Karnin et al. (2013), one can establish the following upper bound on the error probability of Sequential Halving for MTD identification. The proof can be found in Appendix C.

Theorem 3. *The error probability of the* `SH` *algorithm is upper bounded as*

$$\mathbb{P}\left(\hat{k}_n \neq k^*\right) \leq 9 \log_2 K \cdot \exp\left(-\frac{n}{8 H_2(\boldsymbol{p}) \log_2 K}\right),$$

where $H_2(\boldsymbol{p}) := \max_{k \neq k^*} k \Delta_{[k]}^{-2}$ *where* $\Delta_k = |p_k - \theta| - |p_{k^*} - \theta|$ *and* $\Delta_{[1]} \leq \Delta_{[2]} \leq \cdots \leq \Delta_{[K]}$.

A consequence of Theorem 3 is that in a trial involving more than $n = 8 H_2(\boldsymbol{p}) \log_2 K \log\left(9 \log_2(K)/\delta\right)$ patients, Sequential Halving is guaranteed to identify the MTD with probability larger than $1 - \delta$. However,

this number is typically much larger than the number of patients involved in a clinical trial. Indeed the complexity term $H_2(\boldsymbol{p})$ may be quite large, when some doses have a distance to the threshold θ which is very close to the smallest distance $|p_{k^*} - \theta|$.

An important shortcoming of Sequential Halving is that due to the uniform exploration within each phase each dose is selected at least $n/(K \log_2(K))$ times, even the largest, possibly harmful ones. This is highly unethical in a clinical trial without prior knowledge that too-toxic (or too ineffective) doses have already been eliminated. This problem of allocating too extreme doses is likely to be shared by adaptations of any other BAI algorithm, that are expected to select all the arms a linear number of times. For example the APT algorithm proposed by Locatelli et al. (2016) to identify all arms with mean above a threshold θ using a fixed budget n also selects all arms a linear number of times.

To overcome this problem, an interesting avenue of research would be to try to incorporate monotonicity assumptions in BAI algorithms. Garivier et al. (2019a) recently proposed such an algorithm, in the fixed confidence setting: given a risk parameter δ, the goal is to identify a dose $\hat{k}_\tau$ such that $\mathbb{P}(\hat{k}_\tau \neq k^*) \leq \delta$, using as few samples τ as possible. Their analysis identifies a minimal sample complexity $\mathbb{E}[\tau]$ that guarantees a δ-correct identification for any increasing toxicities, which can be obtained under an *optimal allocation* w^* (where w^*_k indicates the fraction of time dose k is allocated). Interestingly, this optimal allocation is supported only on the neighboring doses of the MTD. The fixed-confidence setting requires allowing for random stopping rules τ, i.e. for a dose-finding trial based on an adaptively chosen number of patients. This is not always possible in practice, and it would be interesting to investigate optimal allocations in a fixed-budget setting as well. Yet optimality in the fixed-budget setting is a notoriously hard question already for classical bandits (Carpentier and Locatelli, 2016).

7 Conclusion

Motivated by the literature on multi-armed bandit models, we advocated the use of the powerful Thompson Sampling principle for dose-finding studies. This Bayesian randomized algorithm can be used in different contexts as it can leverage different prior information about the doses. For increasing toxicities and increasing or plateau efficacies, we proposed variants of Thompson Sampling, notably the TS_A algorithm that often outperforms our baselines in terms of recommendation of the optimal dose, while significantly reducing the allocation to doses with high toxicity.

We provided theoretical guarantees for the simplest version of Thompson Sampling based on independent uniform priors on each dose toxicity, but advocated the use of more sophisticated priors for practical dose-finding studies. We believe that finding a practical design for which we can also establish non-trivial finite-time performance guarantees is a crucial research question.

Another interesting direction would be taking contextual information (e.g. a patient's medical history and other medications used) into account for a more "personalized" assessment of toxicity and efficacy of a drug. Bayesian methods also seem promising for such an objective, following the success of Thompson Sampling for contextual bandits.

Acknowledgments

Emilie Kaufmann acknowledges the support of the French Agence Nationale de la Recherche (ANR) under grant ANR-16-CE40-0002 (BADASS project) and ANR-19-CE23-0026-04 (BOLD project). We thank anonymous reviewers of this paper for their helpful suggestions for improvements.

References

Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference On Learning Theory*.

Agrawal, S. and Goyal, N. (2013a). Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.

Agrawal, S. and Goyal, N. (2013b). Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning (ICML)*.

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Anandkumar, A., Michael, N., Tang, A. K., and Agrawal, S. (2011). Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745.

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best Arm Identification in Multi-armed Bandits. In *Proceedings of the 23rd Conference on Learning Theory*.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19).

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256.

Berry, D. (2006). Bayesian clinical trials. *Nature Reviews Drug Discovery*, 5(1).

Berry, S., Carlin, B., Lee, J., and Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC Press.

Brochu, E., Cora, V., and De Freitas, N. (2010). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. Technical report, University of Bristish Columbia.

Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure Exploration in Finitely Armed and Continuous Armed Bandits. *Theoretical Computer Science 412, 1832-1852*, 412:1832–1852.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541.

Carpentier, A. and Locatelli, A. (2016). Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Proceedings of the 29th Conference on Learning Theory (COLT)*.

Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*.

Cheung, Y. and Chappell, R. (2002). A simple technique to evaluate model sensitivity in the continual reassessment method. *Biometrics*, 59:671–674.

Cheung, Y.-K. (2011). *Dose Finding by the Continuous Reassessment Method*. CRC Press.

Chevret, S. (2006). *Statistical Methods for dose-Finding Experiments*. Statistics in Practice. John Wiley and Sons Ltd., Chichester.

Chick, S. E. (2006). Bayesian ideas and discrete event simulation: Why, what and how. In *Proceedings of the 2006 Winter Simulation Conference*.

Faries, D. (1994). Practical modifications of the continual reassessment method for phase I cancer clinical trials. *J Biopharm Stat*, 4(2):147–164.

Food and Drugs Administration (FDA) (2018). Adaptive design clinical trials for drugs and biologics.

Garivier, A., Ménard, P., and Rossi, L. (2019a). Thresholding bandit for dose-ranging: The impact of monotonicity. In *International Conference on Machine Learning, Artificial Intelligence and Applications*.

Garivier, A., Ménard, P., and Stoltz, G. (2019b). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399.

Hoering, A., LeBlanc, M., and Crowley, J. (2011). Seamless phase I-II trial design for assessing toxicity and efficacy for targeted agents. *Clin. Cancer Res.*, 17(4):640–646.

Jacko, P. (2019). The finite-horizon two-armed bandit problem with binary responses: A multidisciplinary survey of the history, state of the art, and myths. *arXiv preprint 1906.10173*.

Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal Exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

Katehakis, M. and Robbins, H. (1995). Sequential choice from several populations. *Proceedings of the National Academy of Science*, 92:8584–8585.

Kaufmann, E., Cappé, O., and Garivier, A. (2012a). On Bayesian Upper-Confidence Bounds for Bandit Problems. In *Proceedings of the 15th conference on Artificial Intelligence and Statistics*.

Kaufmann, E. and Garivier, A. (2017). Learning the distribution with largest mean: two bandit frameworks. *ESAIM: Proceedings and Surveys*, 60:114–131.

Kaufmann, E., Korda, N., and Munos, R. (2012b). Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd conference on Algorithmic Learning Theory*.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvari, C. (2018). Bandit Algorithms. Cambridge University Press.

Le Tourneau, C., Dieras, V., Tresca, P., Cacheux, W., and Paoletti, X. (2010). Current challenges for the early clinical development of anticancer drugs in the era of molecularly targeted agents. *Target Oncol*, 5(1):65–72.

Le Tourneau, C., Gan, H. K., Razak, A. R., and Paoletti, X. (2012). Efficiency of new dose escalation designs in dose-finding phase I trials of molecularly targeted agents. *PLoS ONE*, 7(12):e51039.

Le Tourneau, C., Razak, A. R., Gan, H. K., Pop, S., Dieras, V., Tresca, P., and Paoletti, X. (2011). Heterogeneity in the definition of dose-limiting toxicity in phase I cancer clinical trials of molecularly targeted agents: a review of the literature. *Eur. J. Cancer*, 47(10):1468–1475.

Li, L., Chu, W., Langford, J., and Schapire, R. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th WWW Conference*.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. *arXiv preprint arXiv:1605.08671*.

Mozgunov, P. and Jaki, T. (2017). An information-theoretic approach for selecting arms in clinical trials. *arXiv:1708.02426*.

O'Quigley, J., Pepe, M., and Fisher, L. (1990). Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics*, 46:33–48.

Pallmann, P., Bedding, A., Choodari-Oskooei, B., Dimairo, M., Flight, L., Hampson, L., Holmes, J., Mander, A., Odondi, L., Sydes, M., Villar, S., Wason, J., Weir, C., Wheeler, G., Yap, C., and Jaki, T. (2018). Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(29).

Postel-Vinay, S., Arkenau, H. T., Olmos, D., Ang, J., Barriuso, J., Ashley, S., Banerji, U., De-Bono, J., Judson, I., and Kaye, S. (2009). Clinical benefit in Phase-I trials of novel molecularly targeted agents: does dose matter? *Br. J. Cancer*, 100(9):1373–1378.

Powell, W. B. and Ryzhov, I. O. (2012). *Optimal Learning*. Wiley Series in Probability and Statistics.

Riviere, M.-K., Yuan, Y., Jourdan, J.-H., Dubois, F., and Zohar, S. (2017). Phase i/ii dose-finding design for molecularly targeted agent: Plateau determination using adaptive randomization. *Statistical Methods in Medical Research*.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Satlin, A., Wang, J., Logovinsky, V., Berry, S., Swanson, C., Dhadda, S., and Berry, D. A. (2016). Design of a bayesian adaptive phase 2 proof-of-concept trial for ban2401, a putative disease-modifying monoclonal antibody for the treatment of alzheimer's disease. *Alzheimer's Dementia: Translational Research and Clinical Intervention*, 2(1).

Shen, L. and O'Quigley, J. (1996). Consistency of the continual reassessment method under model misspecification. *Biometrika*, 83:395–405.

Stan Development Team (2015). Stan modeling language users guide and reference manual. *http://mc-stan.org*, version 2.8.0.

Storer, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics*, 45:925–37.

Thall, P. and Wathen, J. (2007). Practical bayesian adaptive randomization in clinical trials. *European Journal on Cancer*, 43:859–866.

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

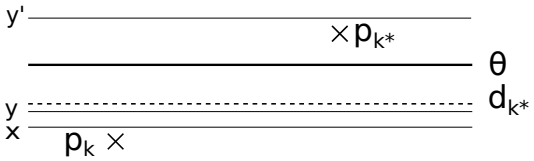
Villar, S., Bowden, J., and Wason, J. (2015). Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges. *Statistical Science*, 30(2):199–215.

Wang, Y., Wang, C., and Powell, W. B. (2016). The knowledge gradient for sequential decision making with stochastic binary feedbacks. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 1138–1147.

Xie, J., Frazier, P. I., and Chick, S. E. (2016). Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs. *Operations Research*, 64(2):542–559.

A Analysis of Independent Thompson Sampling: Proof of Theorem 1

Fix a sub-optimal arm k. Several cases need to be considered depending on the relative position of p_k and p_{k^*} with respect to the threshold. All cases can be treated similarly and to fix the ideas, we consider the case $p_{k^*} \geq \theta > p_k$, which is illustrated below. In that case $d_k^* = 2\theta - p_{k^*}$ satisfies $p_k < d_k^* \leq \theta$.



Let $x, y \in]0,1[^2$ be such that $p_k < x < y < d_{k^*}$, that will be chosen later. Define $y' = 2\theta - y > \theta$ the symmetric of y with respect to the threshold (see the above illustration). We denote by $\hat{\mu}_k(t)$ the empirical mean of the toxicity responses gathered from dose k up to the end of round t and recall $\theta_k(t)$ is the sample from the Beta posterior on p_k after t rounds that is used in the Thompson Sampling algorithm. Inspired by the analysis of Agrawal and Goyal (2013a), we introduce the following two events, that are quite likely to happen when enough samples of arm k have been gathered:

$$E_k^{\mu}(t) \quad = \quad (\hat{\mu}_k(t) \leq x) \quad \text{and} \quad E_k^{\theta}(t) = (\theta_k(t) \leq y)\,.$$

The expected number of allocations of dose k is then decomposed in the following way

$$\mathbb{E}[N_k(T)] = \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left(D_{t+1} = k, E_k^{\mu}(t), E_k^{\theta}(t)\right)}_{(I)} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left(D_{t+1} = k, E_k^{\mu}(t), \overline{E_k^{\theta}(t)}\right)}_{(II)}$$
$$+ \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left(D_{t+1} = k, \overline{E_k^{\mu}(t)}\right)}_{(III)}$$

Terms (II) and (III) are easily controlled using some concentration inequalities and the so-called Beta-Binomial trick, that is the fact that the CDF of a Beta distribution with parameters a and b, $F_{a,b}^{\text{Beta}}$, is related to the CDF of a binomial distribution with parameter n, x, $F_{n,x}^{B}$, in the following way:

$$F_{a,b}^{\text{Beta}}(x) = 1 - F_{a+b-1,x}^{B}(a-1).$$

Term (III) is very small as arm k is unlikely to be drawn often while its empirical mean falls above $x > p_k$ and term (II) grows logarithmically with T. More precisely, it can be shown using Lemma 3 and 4 in Agrawal and Goyal (2013a) that

$$(II) \leq \frac{\log(T)}{\mathrm{kl}(x,y)} + 1 \quad \text{and} \quad (III) \leq \frac{1}{\mathrm{kl}(x,y)} + 1.$$

The tricky part of the analysis is to control term (I), that is to upper bound the number of selections of dose k when both the empirical mean and the Thompson sample for dose k fall close to the true mean p_k. For this purpose, one can prove a counterpart of Lemma 1 in Agrawal and Goyal (2013a) that relates the probability of selecting dose k to that of selecting the MTD k^*.

Lemma 4. *Define $p_y(t) := \mathbb{P}\left(\theta_{k^*}(t) \in [y, y'] | \mathcal{F}_t\right)$, where $\mathcal{F}_s$ is the filtration generated by the observation up to the end of round s. Then*

$$\mathbb{P}\left(D_{t+1} = k | E_k^\theta(t+1), \mathcal{F}_t\right) \leq \frac{1 - p_y(t)}{p_y(t)} \mathbb{P}\left(D_{t+1} = k^* | E_k^\theta(t+1), \mathcal{F}_t\right).$$

Proof. The proof is inspired of that of Lemma 1 in Agrawal and Goyal (2013a). We introduce the event in which the Thompson sample for dose k is the closest to the threshold θ among all sub-optimal doses:

$$M_k(t) = \{|\theta - \theta_k(t)| \geq |\theta - \theta_\ell(t)| \forall \ell \neq k^*\}.$$

On the one hand, one has

$$\begin{aligned}
\mathbb{P}\left(D_{t+1} = k^* | E_k^\theta(t+1), \mathcal{F}_t\right) &\geq \mathbb{P}\left(D_{t+1} = k^*, M_k(t) | E_k^\theta(t+1), \mathcal{F}_t\right) \\
&\geq \mathbb{P}\left(\theta_{k^*}(t) \in [y, y'], M_k(t) | E_k^\theta(t+1), \mathcal{F}_t\right) \\
&= p_y(t) \times \mathbb{P}\left(M_k(t) | E_k^\theta(t+1), \mathcal{F}_t\right).
\end{aligned}$$

On the other hand, it holds that

$$\begin{aligned}
\mathbb{P}\left(D_{t+1} = k | E_k^\theta(t+1), \mathcal{F}_t\right) &\leq \mathbb{P}\left(\theta_{k^*}(t) \notin [y, y'], M_k(t) | E_k^\theta(t+1), \mathcal{F}_t\right) \\
&= (1 - p_y(t)) \times \mathbb{P}\left(M_k(t) | E_k^\theta(t+1), \mathcal{F}_t\right).
\end{aligned}$$

Combining the two inequalities yields Lemma 4.

□

Using the same steps as Agrawal and Goyal (2013a) yields an upper bound on the first term:

$$(I) \leq \sum_{j=1}^{T-1} \mathbb{E}\left[\frac{1}{p_y(\tau_j)} - 1\right],$$

where τ_j is the time instant at which dose k is selected for the j-th time. The expectation of $1/p_y(\tau_j)$ can be explicitly written

$$\mathbb{E}\left[\frac{1}{p_y(\tau_j)}\right] = \sum_{s=0}^{j} \frac{f^B_{j,p_{k^*}}(s)}{\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right)}$$

where $f^B_{n,x}$ stands for the pdf of a Binomial distribution and $X_{a,b}$ denotes a random variable that has a $\mathrm{Beta}(a,b)$ distribution. The following lemma is crucial to finish the proof. This original result was specifically obtained for the MTD identification problem and is needed to control the probability that a Beta distributed random variable fall inside an interval, that is $\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right)$.

Lemma 5. *There exists j_0 such that, for all $j \geq j_0$,*

$$\forall s \in \{0,\dots,j\},\ \ \mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right) \geq \frac{1}{2}\min\left\{\mathbb{P}\left(X_{s+1,j+s+1} \geq y\right); \mathbb{P}\left(X_{s+1,j+s+1} \leq y'\right)\right\}$$

Using Lemma 5 and the Beta-Binomial trick, one can write, for $j \geq j_0$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{p_y(\tau_j)}\right] &\leq \sum_{s=0}^{j} \frac{2f^B_{j,p_{k^*}}(s)}{\mathbb{P}\left(X_{s+1,j+s+1} \geq y\right)} + \sum_{s=0}^{j} \frac{2f^B_{j,p_{k^*}}(s)}{\mathbb{P}\left(X_{s+1,j+s+1} \leq y'\right)} \\
&= \sum_{s=0}^{j} \frac{2f^B_{j,p_{k^*}}(s)}{F^B_{j+1,y}(s)} + \sum_{s=0}^{j} \frac{2f^B_{j,p_{k^*}}(s)}{1-F^B_{j+1,y'}(s)} \\
&= \sum_{s=0}^{j} \frac{2f^B_{j,p_{k^*}}(s)}{F^B_{j+1,y}(s)} + \sum_{s=0}^{j} \frac{2f^B_{j,1-p_{k^*}}(s)}{F^B_{j+1,1-y'}(s)},
\end{aligned} \tag{6}$$

where the last equality relies on the following properties of the Binomial distribution

$$f^B_{n,x}(s) = f^B_{n,1-x}(n-s) \ \text{ and } \ F^B_{n,x}(s) = 1 - F_{n,1-x}(n-s-1)$$

and a change of variable in the second sum.

Now the following upper bound can be extracted from the proof of Lemma 3 in Agrawal and Goyal (2013a).

Lemma 6. *Fix u and v such that $u < v$ and let $\Delta = v - u$. Then*

$$\sum_{s=0}^{j} \frac{f^B_{j,v}(s)}{F^B_{j,u}(s)} \leq \begin{cases} 1 + \frac{3}{\Delta} & \text{if } j < 8/\Delta, \\ 1 + \Theta\left(e^{-\Delta^2 j/2} + \frac{1}{(j+1)\Delta^2} e^{-2\Delta^2 j} + \frac{1}{e^{\Delta^2 j/4}-1}\right) & \text{else.} \end{cases}$$

Each of the two sums in (6) can be upper bounded using Lemma 6. Letting $\Delta_1 = p_{k^*} - y$ and $\Delta_2 = y' - p_{k^*}$, one obtains

$$\begin{aligned}
(I) \leq &\sum_{j=1}^{j_0} \mathbb{E}\left[\frac{1}{p_y(\tau_j)}\right] - j_0 + \frac{24}{\Delta_1^2} + \frac{24}{\Delta_2^2} \\
&+ C\sum_{j=0}^{T-1}\left[e^{-\Delta_1^2 j/2} + \frac{1}{(j+1)\Delta_1^2} e^{-2\Delta_1^2 j} + \frac{1}{e^{\Delta_1^2 j/4}-1}\right] \\
&+ C\sum_{j=0}^{T-1}\left[e^{-\Delta_2^2 j/2} + \frac{1}{(j+1)\Delta_2^2} e^{-2\Delta_2^2 j} + \frac{1}{e^{\Delta_2^2 j/4}-1}\right],
\end{aligned}$$

which is a constant (as the series have a finite sum) that only depends on y, θ and p_{k^*} (through y' and the gaps Δ_1 and Δ_2 defined above).

Putting things together, we proved that for every x and y satisfying $p_k < x < y < d_{k^*}$, the number of selections of dose k is upper bounded as

$$\mathbb{E}[N_k(T)] \leq \frac{1}{\mathrm{kl}(x,y)} \log(T) + C_{x,y,\theta,\boldsymbol{p}}$$

for some constant that depends on the toxicity probabilities, the threshold θ and the choice of x and y. Now, picking x and y such that $\mathrm{kl}(x,y) = \frac{\mathrm{kl}(p_k,d_{k^*})}{1+\epsilon}$ yield the result.

□

Proof of Lemma 5. The proof uses the two equalities below

$$\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right) = \mathbb{P}\left(X_{s+1,j-s+1} \geq y\right) - \mathbb{P}\left(X_{s+1,j-s+1} \geq y'\right) \tag{7}$$
$$\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right) = \mathbb{P}\left(X_{s+1,j-s+1} \leq y'\right) - \mathbb{P}\left(X_{s+1,j-s+1} \leq y\right), \tag{8}$$

as well as the Sanov inequalities: if $S_{n,x}$ is a binomial distribution with parameters n and x, then

$$\begin{aligned} \frac{e^{-n\mathrm{kl}(k/n,x)}}{n+1} &\leq \mathbb{P}\left(S_{n,x} \geq k\right) \\ &\leq e^{-n\mathrm{kl}(k/n,x)} \text{ if } k > xn \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{e^{-n\mathrm{kl}(k/n,x)}}{n+1} &\leq \mathbb{P}\left(S_{n,x} \leq k\right) \\ &\leq e^{-n\mathrm{kl}(k/n,x)} \text{ if } k < xn \end{aligned} \tag{10}$$

We prove the inequality considering 4 cases. We define $y_{\mathrm{mid}} = \frac{y+y'}{2}$.

Case 1: $s < (j+1)y$ Starting from equality (7) and using the Beta-Binomial trick yields

$$\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right) = \mathbb{P}\left(S_{j+1,y} \leq s\right) - \mathbb{P}\left(S_{j+1,y'} \leq s\right).$$

Using Sanov inequalities, we shall prove that there exists some j_1 such that if $j \geq j_1$,

$$\forall s \leq (j+1)y, \;\; \mathbb{P}\left(S_{j+1,y'} \leq s\right) \leq \frac{1}{2}\mathbb{P}\left(S_{j+1,y} \leq s\right).$$

As s is smaller than the mean of the two Binomial distributions, by (10) it is sufficient to prove that

$$\forall s \leq (j+1)y, \;\; e^{-(j+1)\mathrm{kl}\left(\frac{s}{j+1},y'\right)} \leq \frac{1}{2(j+2)} e^{-(j+1)\mathrm{kl}\left(\frac{s}{j+1},y\right)}$$

which in turn is equivalent to

$$\forall s \leq (j+1)y, \;\; \mathrm{kl}\left(\frac{s}{j+1},y'\right) - \mathrm{kl}\left(\frac{s}{j+1},y\right) \geq \frac{\log(2(j+2))}{j+1}.$$

As the function in the left-hand side is non-increasing in s, a sufficient condition is that j satisfies

$$\mathrm{kl}\left(y,y'\right) \geq \frac{\log(2(j+2))}{j+1},$$

which is the case for j superior to some j_1. Thus, for $j \geq j_1$,

$$\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right) \geq \frac{1}{2}\mathbb{P}\left(S_{j+1,y} \leq s\right) = \frac{1}{2}\mathbb{P}\left(X_{s+1,j-s+1} \geq y\right).$$

Case 2: $(j+1)y \le s \le (j+1)y_{\mathrm{mid}}$ Starting from equality (7) and using the Beta-Binomial trick and the upper bound in (10) yields

$$\begin{aligned}\mathbb{P}\left(y \le X_{s+1,j-s+1} \le y'\right) &\ge \mathbb{P}\left(S_{j+1,y} \le s\right) - e^{-(j+1)\mathrm{kl}\left(\frac{s}{j+1},y'\right)} \\ &\ge \mathbb{P}\left(S_{j+1,y} \le s\right) - e^{-(j+1)\mathrm{kl}(y_{\mathrm{mid}},y')}.\end{aligned}$$

The median of $S_{j+1,y}$ is $\lfloor (j+1)y \rfloor$ or $\lceil (j+1)y \rceil$. As $s \le (j+1)y$, it holds that $\mathbb{P}\left(S_{j+1,y} \le s\right) \ge \frac{1}{2}$. Therefore, for all $j \ge j_2 := \frac{\ln 4}{\mathrm{kl}(y_{\mathrm{mid}},y')} - 1$,

$$e^{-(j+1)\mathrm{kl}(y_{\mathrm{mid}},y')} \le \frac{1}{4} \le \frac{1}{2}\mathbb{P}\left(S_{j+1,y} \le s\right).$$

Therefore if $j \ge j_2$, $\mathbb{P}\left(y \le X_{s+1,j-s+1} \le y'\right) \ge \frac{1}{2}\mathbb{P}\left(X_{s+1,j-s+1} \ge y\right)$.

Case 3: $(j+1)y_{\mathrm{mid}} \le s \le (j+1)y'$ Starting from equality (8) and using the Beta-Binomial trick and the upper bound in (9) yields

$$\begin{aligned}\mathbb{P}\left(y \le X_{s+1,j-s+1} \le y'\right) &\ge \mathbb{P}\left(S_{j+1,y'} \ge s\right) - e^{-(j+1)\mathrm{kl}\left(\frac{s}{j+1},y\right)} \\ &\ge \mathbb{P}\left(S_{j+1,y'} \ge s\right) - e^{-(j+1)\mathrm{kl}(y_{\mathrm{mid}},y)}.\end{aligned}$$

The median of $S_{j+1,y'}$ is $\lfloor (j+1)y' \rfloor$ or $\lceil (j+1)y' \rceil$. As $s \le (j+1)y'$, it holds that $\mathbb{P}\left(S_{j+1,y'} \ge s\right) \ge \frac{1}{2}$. Therefore, for all $j \ge j_3 := \frac{\ln 4}{\mathrm{kl}(y_{\mathrm{mid}},y)} - 1$,

$$e^{-(j+1)\mathrm{kl}(y_{\mathrm{mid}},y)} \le \frac{1}{4} \le \frac{1}{2}\mathbb{P}\left(S_{j+1,y'} \ge s\right).$$

Therefore if $j \ge j_3$, $\mathbb{P}\left(y \le X_{s+1,j-s+1} \le y'\right) \ge \frac{1}{2}\mathbb{P}\left(X_{s+1,j-s+1} \le y'\right)$.

Case 4: $s > (j+1)y'$ Starting from equality (8) and using the Beta-Binomial trick yields

$$\mathbb{P}\left(y \le X_{s+1,j-s+1} \le y'\right) = \mathbb{P}\left(S_{j+1,y'} \ge s\right) - \mathbb{P}\left(S_{j+1,y} \ge s\right).$$

Using Sanov inequalities, we shall prove that there exists some j_4 such that if $j \ge j_4$,

$$\forall s \ge (j+1)y', \;\; \mathbb{P}\left(S_{j+1,y} \ge s\right) \le \frac{1}{2}\mathbb{P}\left(S_{j+1,y'} \ge s\right).$$

As s is larger than the mean of the two Binomial distributions, by (9) it is sufficient to prove that

$$\forall s \ge (j+1)y', \;\; e^{-(j+1)\mathrm{kl}\left(\frac{s}{j+1},y\right)} \le \frac{1}{2(j+2)}e^{-(j+1)\mathrm{kl}\left(\frac{s}{j+1},y'\right)}$$

which in turn is equivalent to

$$\forall s \ge (j+1)y', \;\; \mathrm{kl}\left(\frac{s}{j+1},y\right) - \mathrm{kl}\left(\frac{s}{j+1},y'\right) \ge \frac{\log(2(j+2))}{j+1}.$$

As the function in the left-hand side is non-decreasing in s, a sufficient condition is that j satisfies

$$\mathrm{kl}\left(y',y\right) \ge \frac{\log(2(j+2))}{j+1},$$

which is the case for j superior to some j_4. Thus, for $j \geq j_4$,

$$\begin{aligned}\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right) &\geq \frac{1}{2}\mathbb{P}\left(S_{j+1,y'} \geq s\right) \\ &= \frac{1}{2}\mathbb{P}\left(X_{s+1,j-s+1} \leq y'\right).\end{aligned}$$

Conclusion Letting $j_0 = \max(j_1, j_2, j_3, j_4)$, for all $j \geq j_0$, for every $s \in \{0, \ldots, j\}$,

$$\mathbb{P}\left(y \leq X_{s+1,j-s+1} \leq y'\right) \geq \frac{1}{2}\min\left\{\mathbb{P}\left(X_{s+1,j+s+1} \geq y\right); \mathbb{P}\left(X_{s+1,j+s+1} \leq y'\right)\right\}$$

B Lower Bound on the Number of Allocation: Proof of Theorem 2

Fix a uniformly efficient algorithm and a vector of toxicity probabilities $\boldsymbol{p}$. We denote by $\mathbb{E}_{\boldsymbol{p}}$ the expectation under the model parameterized by $\boldsymbol{p}$ when this algorithm is used. Letting $\boldsymbol{p}'$ be another vector of probabilities, it follows from the change-of-distribution lemma of Garivier et al. (2019b) that for all random variable $Z_T \in [0,1]$ which is $\mathcal{F}_T$-measurable

$$\sum_{\ell=1}^{K} \mathbb{E}_{\boldsymbol{p}}[N_\ell(T)]\mathrm{kl}\left(p_\ell, p'_\ell\right) \geq \mathrm{kl}\left(\mathbb{E}_{\boldsymbol{p}}[Z_T], \mathbb{E}_{\boldsymbol{p}'}[Z_T]\right). \tag{11}$$

Letting k^* be a MTD in $\boldsymbol{p}$, we fix k which is not a MTD (i.e. $|p_k - \theta| > |p_{k^*} - \theta|$) and we prove that

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{p}}[N_k(T)]}{\ln(T)} \geq \frac{1}{\mathrm{kl}(p_k, d_k^*)}. \tag{12}$$

Recall that we assume $p_{k^*} \neq \theta$. Then one can define the alternative model $\boldsymbol{p}'$ in which for all $\ell \neq k$, $p'_\ell = p_\ell$ and $p'_k = d_k^* + \epsilon$ if $d_k^* < \theta$ and $p'_k = d_k^* - \epsilon$ if $d_k^* > \theta$, with ϵ small enough such that under $\boldsymbol{p}'$, dose k is the unique MTD (refer to Figure 1 for an illustration).

For this particular choice of alternative model $\boldsymbol{p}'$, (11) becomes

$$\begin{aligned}\mathbb{E}_{\boldsymbol{p}}[N_k(T)]\mathrm{kl}(p_k, d_k^* \pm \epsilon) &\geq \mathrm{kl}\left(\mathbb{E}_{\boldsymbol{p}}[Z_T], \mathbb{E}_{\boldsymbol{p}'}[Z_T]\right) \\ &\geq (1 - \mathbb{E}_{\boldsymbol{p}}[Z_T])\ln\left(\frac{1}{1 - \mathbb{E}_{\boldsymbol{p}'}[Z_T]}\right) - \ln(2)\end{aligned}$$

Choosing $Z_T = \frac{N_k(T)}{T}$, exploiting the fact that the algorithm is uniformly efficient we know that

- $\lim_{T\to\infty} \mathbb{E}_{\boldsymbol{p}}[Z_T] = 0$ as k is a sub-optimal dose under $\boldsymbol{p}$
- $\frac{1}{1-\mathbb{E}_{\boldsymbol{p}'}[Z_T]} = \frac{T}{T-\mathbb{E}_{\boldsymbol{p}'}[N_k(T)]} = \frac{T}{\sum_{\ell\neq k}\mathbb{E}_{\boldsymbol{p}'}[N_\ell(T)]}$ and $\sum_{\ell\neq k}\mathbb{E}_{\boldsymbol{p}'}[N_\ell(T)] = o(T^\alpha)$ for all $\alpha \in (0,1)$ as k is the only MTD under $\boldsymbol{p}'$, which yields, for all $\alpha \in (0,1)$,

$$\lim_{T\to\infty} \frac{1}{\ln(T)} \ln\left(\frac{1}{1 - \mathbb{E}_{\boldsymbol{p}'}[Z_T]}\right) \geq 1 - \alpha.$$

Letting α go to zero, we obtain

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{p}}[N_k(T)]\mathrm{kl}(p_k, d_k^* \pm \epsilon)}{\ln(T)} \geq 1$$

and (12) follows by letting ϵ go to zero.

C Analysis of Sequential Halving: Proof of Theorem 3

Recall $\hat{d}_k^r = |\theta - \hat{p}_k^t|$ is the empirical distance from the toxicity of dose k to the threshold, where $\hat{p}_k^r$ is the empirical average of the toxicity responses observed for dose k during phase r (based on t_r samples). The central element of the proof is Lemma 7 below, that controls the probability that dose k seems to be be closer to the threshold than the MTD k^* in phase r. Its proof is more sophisticated than that of Lemma 4.2 in Karnin et al. (2013) as several cases need to be considered.

Lemma 7. *Assume that the arm closest to θ was not eliminated prior to round r. Then for any arm $k \in S_r$,*

$$\mathbb{P}(\hat{d}_{k^*}^r > \hat{d}_k^r) \leq 3\exp\left(-\frac{t_r}{2}\Delta_k^2\right). \tag{13}$$

Proof. For the means p_{k^*} and p_k let $\hat{p}_{k^*}^r$ and $\hat{p}_k^r$ denote their expected rewards in round r, respectively. We will first derive a probability bound which does not depend on the ordering of p_k and p_{k^*} w.r.t. θ, and then we will do a case analysis of the possible orderings to produce our final bound.

The error event can be decomposed as follows.

$$\begin{aligned}
&\left\{\hat{d}_{k^*}^r > \hat{d}_k^r\right\} = \\
&\quad(\{\hat{p}_{k^*,r} > \theta\} \cap \{\hat{p}_{k,r} > \theta\} \cap \{\hat{p}_{k^*,r} - \theta > \hat{p}_{k,r} - \theta\}) \\
&\cup(\{\hat{p}_{k^*,r} \leq \theta\} \cap \{\hat{p}_{k,r} > \theta\} \cap \{\theta - \hat{p}_{k^*,r} > \hat{p}_{k,r} - \theta\}) \\
&\cup(\{\hat{p}_{k^*,r} > \theta\} \cap \{\hat{p}_{k,r} \leq \theta\} \cap \{\hat{p}_{k^*,r} - \theta > \theta - \hat{p}_{k,r}\}) \\
&\cup(\{\hat{p}_{k^*,r} \leq \theta\} \cap \{\hat{p}_{k,r} \leq \theta\} \cap \{\theta - \hat{p}_{k^*,r} > \theta - \hat{p}_{k,r}\})
\end{aligned}$$

From there, we distinguish two cases, in which we show the error event is included in a reunion of events whose probability can be controlled using the Hoeffding's inequality.

Case 1: $p_k \geq \theta$. In that case, it is very unlikely that $\{\hat{p}_{k,r} < \theta\}$. Hence, we can isolate that event and use the previous decomposition to write

$$\begin{aligned}
&\left\{\hat{d}_{k^*}^r > \hat{d}_k^r\right\} \subseteq \\
&\{\hat{p}_{k,r} \leq \theta\} \cup \{\hat{p}_{k^*,r} - \hat{p}_{k,r} > 0\} \cup \{\hat{p}_{k,r} + \hat{p}_{k^*,r} < 2\theta\}.
\end{aligned}$$

When $p_k \geq \theta$, irrespective of the position of p_{k^*} with respect to θ, one can justify that $p_k > \theta$, $p_{k^*} - p_k < 0$ and $p_k + p_{k^*} > 2\theta$ (as $p_k \geq \max(p_{k^*}, 2\theta - p_{k^*})$ because k is a suboptimal arm larger than the threshold). Therefore, the above three events are unlikely. More precisely, using Hoeffding's inequality yields

$$\begin{aligned}
\mathbb{P}(\hat{d}_{k^*}^r > \hat{d}_k^r) &\leq \mathbb{P}(\hat{p}_{k,r} \leq \theta) + \mathbb{P}(\hat{p}_{k^*,r} - \hat{p}_{k,r} > 0) + \mathbb{P}(\hat{p}_{k^*,r} + p_{k,r} < 2\theta) \\
&\leq \exp\left(-2t_r(\theta - p_k)^2\right\} + \exp\left\{-\frac{t_r}{2}(p_{k^*} - p_k)^2\right\} \\
&\quad + \exp\left\{-\frac{t_r}{2}(p_{k^*} + p_k - 2\theta)^2\right) \\
&\leq 3\exp\left(-\frac{t_r}{2}\min\left\{(p_k - \theta)^2, (p_k - p_{k^*})^2, (p_{k^*} + p_k - 2\theta)^2\right\}\right) \\
&= 3\exp\left(-\frac{t_r}{2}\min\left\{(p_k - p_{k^*})^2, (p_k - (2\theta - p_{k^*}))^2\right\}\right)
\end{aligned}$$

Equation (13) follows as $\Delta_k^2 = \min\left\{(p_k - p_{k^*})^2, (p_k - (2\theta - p_{k^*}))^2\right\}$.

Case 2: $p_k \leq \theta$. In that case, the unlikely event is $\{\hat{p}_{k,r} > \theta\}$ and we write

$$\left\{\hat{d}^r_{k^*} > \hat{d}^r_k\right\} \subseteq \{\hat{p}_{k,r} > \theta\} \cup \{\hat{p}_{k,r} - \hat{p}_{k^*,r} > 0\} \cup \{\hat{p}_{k,r} + \hat{p}_{k^*,r} > 2\theta\}\,.$$

When $p_k < \theta$, irrespective of the position of p_{k^*} with respect to θ, one can justify that $p_k < \theta$, $p_k - p_{k^*} < 0$ and $p_k + p_{k^*} < 2\theta$ (using the fact that $p_k \leq \min(p_{k^*}, 2\theta - p_{k^*})$). Then from Hoeffding's inequality,

$$\begin{aligned}
\mathbb{P}(\hat{d}^r_{k^*} > \hat{d}^r_k) &\leq \mathbb{P}(\hat{p}_{k,r} > \theta) + \mathbb{P}(\hat{p}_{k,r} - \hat{p}_{k^*,r} > 0) + \mathbb{P}(\hat{p}_{k^*,r} + p_{k,r} > 2\theta) \\
&\leq \exp\left(-2t_r(\theta - p_k)^2\right\} + \exp\left\{-\frac{t_r}{2}(p_{k^*} - p_k)^2\right\} \\
&\quad + \exp\left\{-\frac{t_r}{2}(2\theta - p_{k^*} - p_k)^2\right) \\
&\leq 3\exp\left(-\frac{t_r}{2}\min\left\{(\theta - p_k)^2, (p_{k^*} - p_k)^2, (2\theta - p_{k^*} - p_k)^2\right\}\right) \\
&= 3\exp\left(-\frac{t_r}{2}\min\left\{(p_{k^*} - p_k)^2, ((2\theta - p_{k^*}) - p_k)^2\right\}\right)
\end{aligned}$$

which proves Equation 13 as $\Delta_k^2 = \min\left\{(p_{k^*} - p_k)^2, ((2\theta - p_{k^*}) - p_k)^2\right\}$.

□

Building on Lemma 7, the next step is to control the probability that the MTD is eliminated in phase r. The proof bears strong similarities with that of Lemma 4.3 in Karnin et al. (2013). It is given below for the sake of completeness.

Lemma 8. *The probability that the MTD is eliminated at the end of phase r is at most*

$$9\exp\left(-\frac{n}{8\log_2 K}\cdot\frac{\Delta^2_{k_r}}{k_r}\right)$$

where $k_r = K/2^{r+2}$.

The end of the proof of Theorem 3 is identical to than of Theorem 4.1 in Karnin et al. (2013), except that it uses our Lemma 8. We repeat the argument below with the appropriate modifications. Observe that if the algorithm recommends a wrong dose, the MTD must have been eliminated in one of t $\log_2(K)$ phases. Using Lemma 8 and a union bound yields the upper bound

$$\begin{aligned}
\mathbb{P}\left(\hat{k}_n \neq k^*\right) &\leq 9\sum_{r=1}^{\log_2 K}\exp\left(-\frac{n}{8\log_2 K}\cdot\frac{\Delta^2_{k_r}}{k_r}\right) \\
&\leq 9\log_2 K\cdot\exp\left(-\frac{n}{8\log_2 K}\cdot\frac{1}{\max_k k\Delta_k^{-2}}\right) \\
&\leq 9\log_2 K\cdot\exp\left(-\frac{n}{8H_2(\boldsymbol{p})\log_2 K}\right),
\end{aligned}$$

which concludes the proof.

Proof of Lemma 8 Define S'_r as the set of arms in S_r, excluding the $\frac{1}{4}|S_r| = K/2^{r+2}$ arms with means closest to θ. If the MTD k^* is eliminated in round r, it must be the case that at least half the arms of S_r (i.e., $\frac{1}{2}|S_r| = K/2^{r+1}$ arms) have their empirical average closer to θ than its empirical average. In particular, the empirical means of at least $\frac{1}{3}|S'_r| = K/2^{r+2}$ of the arms in S'_r must be closer to θ than that of the k^* at the end of round r. Letting N_r denote the number of arms in S'_r whose empirical average is closer to θ than that of the optimal arm, we have by Lemma 7:

$$\begin{aligned}
\mathbb{E}[N_r] &= \sum_{k \in S'_r} \mathbb{P}(\hat{d}^r_k < \hat{d}^r_{k^*}) \\
&\leq \sum_{k \in S'_r} 3 \exp\left(-\frac{t_r}{2}\Delta_k^2\right) \\
&\leq 3 \sum_{k \in S'_r} \exp\left(-\frac{1}{2}\Delta_k^2 \cdot \frac{n}{|S_r| \log_2 K}\right) \\
&\leq 3|S'_r| \max_{k \in S'_r} \exp\left(-\frac{1}{2}\Delta_k^2 \cdot \frac{2^r n}{K \log_2 K}\right) \\
&\leq 3|S'_r| \exp\left(-\frac{n}{8 \log_2 K} \cdot \frac{\Delta_{k_r}^2}{k_r}\right)
\end{aligned}$$

Where the last inequality follows from the fact that there are at least $k_r - 1$ arms that are not in S'_r with average reward closer to θ than that of any arm in S'_r. We now apply Markov's inequality to obtain

$$\begin{aligned}
\mathbb{P}\left(N_r > \frac{1}{3}|S'_r|\right) &\leq 3\mathbb{E}[N_r]/|S'_r| \\
&\leq 9 \exp\left(-\frac{n}{8 \log_2 K} \cdot \frac{\Delta_{k_r}^2}{k_r}\right),
\end{aligned}$$

and the lemma follows.