Boston University

OpenBU **http://open.bu.edu**

BU Open Access Articles BU Open Access Articles

2022-07-06

Strategies for safe multi-armed bandits with logarithmic regret and risk

V. Saligrama, A. Gangrade, T. Chen. "Strategies for Safe Multi-Armed Bandits with Logarithmic Regret and Risk" International Conference on Machine Learning.

https://hdl.handle.net/2144/47054

Downloaded from DSpace Repository, DSpace Institution's institutional repository

arXiv:2204.00706v1 [cs.LG] 1 Apr 2022

Strategies for Safe Multi-Armed Bandits with Logarithmic Regret and Risk

Tianrui Chen
Boston University
trchen@bu.edu

Aditya Gangrade
Carnegie Mellon University*
agangra2@andrew.cmu.edu

Venkatesh Saligrama
Boston University
srv@bu.edu

Abstract

We investigate a natural but surprisingly unstudied approach to the multi-armed bandit problem under safety risk constraints. Each arm is associated with an unknown law on safety risks and rewards, and the learner's goal is to maximise reward whilst not playing unsafe arms, as determined by a given threshold on the mean risk.

We formulate a pseudo-regret for this setting that enforces this safety constraint in a per-round way by softly penalising any violation, regardless of the gain in reward due to the same. This has practical relevance to scenarios such as clinical trials, where one must maintain safety for each round rather than in an aggregated sense.

We describe doubly optimistic strategies for this scenario, which maintain optimistic indices for both safety risk and reward. We show that schema based on both frequentist and Bayesian indices satisfy tight gap-dependent logarithmic regret bounds, and further that these play unsafe arms only logarithmically many times in total. This theoretical analysis is complemented by simulation studies demonstrating the effectiveness of the proposed schema, and probing the domains in which their use is appropriate.

1 Introduction

We consider the safety constrained multi-armed bandit problem, where each *arm*, $k \in [1:K]$ is modelled by a tuple, consisting of a stochastic *reward*, of mean μ^k, and an associated stochastic *safety-risk*, of mean ν^k. Upon playing an arm, the learner observes noisy instances of the reward and safety-risk. The learner is provided with a *tolerated risk level*, denoted α, and the goal of the *safe bandit problem* is to maximise the reward gained over the course of play, while ensuring that unsafe arms—those for which $\nu^k > \alpha$—are not played too often.

We propose the following *regret* formulation to model the above criteria. Let μ^* be the mean reward of the largest safe action, i.e, the largest μ^k over arms such that $\nu^k \leq \alpha$. Let A_t be the arm pulled by the algorithm at time t. We study

$$\mathcal{R}_T := \sum_{t \leq T} \max(\mu^* - \mu^{A_t}, \nu^{A_t} - \alpha). \tag{1}$$

Before describing the results, let us sketch a scenario of particular interest, which informs our formulation.

Clinical Trials. Trial drugs have both positive (eg. curing a disease) and negative side-effects (headaches, nausea, etc) on a patient in a clinical trial, and it is as much in the interest of a patient to ensure that negative side effects are limited as it is to ensure that the drug is effective (e.g. Genovese et al., 2013). This scenario motivates the problem of choosing drug and dosage (arms) that have the maximum positive response while ensuring that the side-effects remain below some threshold α. Since each patient responds differently, the observed response and the manifestation of side-effects for a specific patient can be modelled as random-variables, with the corresponding means representing population averages. Importantly, for such a scenario, safety must be accounted for in a per-round sense - it does no good to alternate between assigning ineffective

*The bulk of this work was done whilst A.G. was a graduate student at Boston University

placebos and effective but harmful doses. Instead we need to ensure that individuals are not exposed to undue risk while accruing benefits.

How does our formulation account for this scenario?

- *Risk Per Round.* Regret ensures that unsafe arms are rarely played in a per-round (per-patient) sense rather than ensuring safety in an overall sense–for any $k \neq k^*$, at least one of $\mu^k - \mu^*$ or $\nu^k - \alpha$ must be positive, and so benefits in efficacy due to unsafe dosages are discounted.
- *Small safety violations are penalized less (smoothness).* Small violations of negative side-effects is a permissible risk (elevated nausea level than desired), worth taking on for a few patients, in the hope of finding a drug/dosage that is effective for the population. Our penalty on safety violations is smooth.
- *Control of Cumulative risk and Violations* Since choosing an infeasible arm in any round contributes a constant amount to the regret, a small $\mathcal{R}_T$ further ensures that the cumulative safety risk and the cumulative safety violations (i.e. times such that $\nu^{A_t} > \alpha$) are also small.

We next describe our main technical contributions.

Four Optimistic Strategies. We explore *doubly optimistic* index-based strategies for choosing arms. These maintain optimistic indices for both the reward and safety risk of each arm, and proceed by first developing a set of plausibly safe actions using the safety indices, and then choose the arm with the highest reward index to play, thus encouraging sufficient exploration. In standard bandits there are two broad classes of such index-based strategies - those based on *frequentist confidence bounds*, and those based on *Bayesian posteriors*. This suggests four natural variants in the safe bandit case, through two choices for each of the reward and safety indices. We explicitly study three of these - first when both indices are frequentist, second when the safety index is left frequentist but the reward index is replaced by *Thompson sampling*, and finally when both indices are based on Bayesian methods. While left explicitly unstudied, the case of frequentist reward and Bayesian safety indices follows naturally from our analysis.

Logarithmic Regret Bounds. In all cases, we show that these strategies admit strong gap-dependent logarithmic regret rates. Further each of these also ensure that the number of times any unsafe arm is played at all (i.e., $\sum \mathbb{1}\{\nu^{A_t} > \alpha\}$) is similarly logarithmically bounded. Finally we show a lower bound which demonstrates that our regret bounds are tight in the limit of large time horizons. The proofs adapt existing results of bandit theory to argue that for well designed safety indices, the optimal arm k^* always remains valid, but any unsafe arms are quickly eliminated. Further, so long as k^* remains valid, standard approaches show that inefficient arms cannot be played too often. *An interesting consequence is that the play of strictly dominated arms - those that are both unsafe and inefficient - is limited by the larger of the two gaps.*

Empirical Results. We complement the above theoretical study with simulations. First, we practically illustrate that prior policy-based approaches to the safe and constrained bandits do not yield favourable play in our scenario. Next, we implement our proposals, and both illustrate that the methods indeed meet the theoretical guarantees, and further contextualise their relative merits in a practical sense. The broad observation regarding the latter is that Thompson sampling based methods tend to offer better performance in terms of means.

1.1 Related Work

Bandit problems are exceedingly well studied, and a plethora of methods with subtle differences have been established. We refer the reader to the recent book of Lattimore & Szepesvári (2020) for a thorough introduction.

We first describe prior approaches to constrained bandit problems from a formulational point of view. The most important aspect of this is that prior formulations tend to constrain play in an aggregate sense. This raises issues when we need to ensure safety in a per-round sense, as is illustrated by a running example. We then contextualise our methodological proposals with respect to the prior work, and finally discuss pure exploration in the safe-bandit setting.

Globally Constrained Formulations The theory of bandits with global constraints was initiated by Badanidiyuru et al. (2013), and extended by Agrawal & Devanur (2014). Specialised to our context, these works constrain the total number of adverse effects whilst matching the performance of the optimal dynamic policy that is aware of all means. More concretely, suppose that the safety risk observed is a random variable S_t. Badanidiyuru et al. (2013) enforce the hard constraint that $(\sum S_t - \alpha T) \leq 0$, while Agrawal & Devanur (2014) relax this into a second regret $\mathcal{S}_T = \max(0, \sum S_t - \alpha T)$, and ensure that this is small.

Such aggregate safety formulation is lacking from our perspective, as is illustrated by the following simple example of two arms with means

$$(\mu^1, \nu^1) = (1/2, 0), \quad (\mu^2, \nu^2) = (1, 1). \tag{2}$$

Due to the global constraint, the optimal dynamic policy is to pull arm 2 for αT rounds, and then switch to pulling arm 1. A low regret algorithm must then also pull arm 2 $\Omega(T)$ times. However, such play undesirably exposes a linear number of rounds to the very unsafe action 2. Our formulation instead would penalise every play of arm 2 by a cost of $(1 - \alpha)$, and thus effective schema would only play arm 2 sublinearly many times. It should be noted that since the constraint is applied in a per-round way, the optimal dynamic policy in our case is supported on a single arm.

In passing, we also mention the conservative bandit problem (Wu et al., 2016), which only considers rewards, and enforces a running aggregate constraint that for any round t, $\sum_{s \leq t} \mu^{A_s} \geq (1 - \alpha) t \mu^{k_0}$. While an interesting variation, we note that such a running constraint on safety-risk would have similar issues as the above in our situation.

Per-round Constraints The recent work of Pacchiano et al. (2021) studies the safe bandit problem with two crucial differences from us. Firstly, the action space is lifted from single arms to policies (i.e. distributions) over arms, denoted π_t, and secondly, the hard per-round constraint $\langle \pi_t, \nu \rangle \leq \alpha$ is enforced. Of course, actual arms are selected by sampling from π_t. The regret studied is $\sum \langle \pi^* - \pi_t, \mu \rangle$, where π^* is the optimal static safe policy, i.e., the maximiser of $\langle \pi, \mu \rangle$ subject to $\langle \pi, \nu \rangle \leq \alpha$. Exploration is enabled by giving the scheme an arm k_s known a priori to be safe, and by spending the slack $\alpha - \nu^{k_s}$ as room for exploration in π_t.

While ostensibly constrained at each round, this formulation suffers from similar issues as the previously discussed globally constrained formulations since the optimal static policy is only safe in aggregate. Indeed, in the previous example (2), the optimal π^* is $(1 - \alpha, \alpha)$, and so a low regret algorithm must place large mass on the unsafe arm 2 in most rounds, therefore exposing about $\Omega(T)$ rounds to it.

A similar approach, but crucially without the policy action space, was taken by Amani et al. (2019); Moradipari et al. (2021) for in the linear bandit setting. These papers also study hard round-wise safety constraints, and again utilise a known safe action, as well as the continuity of the action space to enable sufficient exploration. We note that the particulars of the signalling model adopted by Amani et al. (2019) paper preclude extending their results to the multi-armed setting, and while the model of Moradipari et al. (2021) does admit such extension, the scheme proposed fundamentally relies on having a continuous action space with a linear safety-risk, and cannot be extended to multi-armed settings without lifting to policy space.

Methodological Approaches The bulk of the previous papers are based on frequentist confidence bounds, with two variants. Similar to our Alg. 1, Agrawal & Devanur (2014) use doubly optimistic methods that maintain optimistic upper bounds on the rewards and lower bounds on the risk, and play the policy that maximises reward upper bounds while being safe with respect to the risk lower bounds. In contrast, Pacchiano et al. (2021); Amani et al. (2019); Wu et al. (2016) all use optimistic-pessimistic methods, which instead maintain upper bounds on both the rewards and safety risk and play the actions with maximum reward upper bound whilst being safe with respect to the stringent risk upper bounds. Moradipari et al. (2021) take a similar pessimistic approach, but replace the reward upper bounds with a Thompson sampling procedure that is similar in spirit to our Alg. 2, although this uses optimistic safety indices. We also further study a fully Bayesian approach in Alg. 3.

Pure Exploration with Safety Katz-Samuels & Scott (2018, 2019) design procedures for finding the best *feasible* arm based on a combination of optimistic and pessimistic confidence bounds that is typical of pure exploration approaches. An interesting variant of this problem was studied in a recent preprint of Wang et al. (2021), who associate a continuous 'dosage' parameter with each arm, now interpreted as a single drug, with the understanding that both reward and risk grow monotonically with dosage. These should be compared to the dose-finding bandit problem Aziz et al. (2021), which seeks to identify a dose level out of K options that minimises $|\nu^k - \alpha|$, with the intuition being that higher doses are more effective, and so should be maximised, but without exceeding the safety threshold by much. The dose-finding approach relies strongly on this assumed monotonicity. This models the scenario of a single drug, but is inappropriate for the setting of multiple drugs that are trialled together, which is better represented as a constrained optimisation problem (as studied by the former papers). Our formulation takes precisely this view, but from the perspective of controlling regret rather than identification. Note that our smooth penalty for safety violation, $\max(0, \nu^k - \alpha)$, bears similarities to the absolute value loss $|\nu^k - \alpha|$, where again a small violation of safety is not penalised strongly.

2 Definitions and Setup

An instance of the *safe bandit problem* is defined by a risk level $\alpha \in [0, 1]$, a natural $K \geq 2$, corresponding to a number of arms, and a corresponding vector of probability distributions, $(\mathbb{P}^k)_{k\in[1:K]}$, each entry of which is supported on $[0, 1]^2$. We will represent the corresponding random vector as two components (R, S), which are termed the reward and safety-risk of a draw from $\mathbb{P}^k$. We further associate two vectors $\mu, \nu \in [0, 1]^K$, corresponding to the *mean reward and safety-risk* of each arm, i.e

$$(\mu^k, \nu^k) := \mathbb{E}_{(R,S)\sim\mathbb{P}^k}[(R, S)].$$

R and S need not be independent - this has little effect on the subsequent study, since each is marginally bounded.

The scenario proceeds in rounds, denoted $t \in \mathbb{N}$. At each t, the learner (i.e. an algorithm for the bandit problem) must choose an *action* $A_t \in [1 : K]$, corresponding to 'pulling an arm.' Upon doing so, the learner receives samples $(R_t, S_t) \sim \mathbb{P}^{A_t}$ independently of the history. The learner's *information set* at time t is $\mathscr{H}_{t-1} = \{(A_s, R_s, S_s) : s < t\}$, and the action A_t must be adapted to the filtration induced by these sets. The learner is unaware of any properties of the laws $\mathbb{P}^k$ beyond the fact that they are supported on $[0, 1]^2$.

The *competitor*, representing the *best safe arm* given the safety constraint and the mean vectors, is defined as

$$k^* = \operatorname*{arg\,max}_{k\in[1:K]} \mu^k \text{ s.t. } \nu^k \leq \alpha,$$

and its mean reward and safety risk are denoted as μ^*, ν^*. We will use this convention throughout - for any symbol $\mathfrak{s}^k$, we set $\mathfrak{s}^* = \mathfrak{s}^{k^*}$. We can ensure that the problem is feasible by including a no-reward, no-risk arm of means $(0, 0)$ - this might correspond to a placebo in a clinical trial. Without loss of generality, we will assume that k^* is unique. We define the *inefficiency gap* Δ^k and the *safety gap* Γ^k of playing an arm k as

$$\Delta^k := 0 \vee (\mu^* - \mu^k), \quad \Gamma^k := 0 \vee (\nu^k - \alpha),$$

where $a \vee b := \max(a, b)$, and we will also use $a \wedge b := \min(a, b)$. Note that $\Delta^k \vee \Gamma^k > 0$ for $k \neq k^*$.

The performance of a learner for the safe bandit problem is measured by the (pseudo-) *regret* of (1), which may also be written as $\mathcal{R}_T := \sum_{1\leq t\leq T} \Delta_{A_t} \vee \Gamma_{A_t}$.

Further, with each arm k, we associate state variables N_t^k denoting the number of times it has been played

up to time t, and R_t^k, S_t^k denoting the total rewards and safety risk incurred on such rounds. More formally,

$$N_t^k := \sum_{s<t} \mathbb{1}\{A_t = k\},$$
$$R_t^k := \sum_{s<t} \mathbb{1}\{A_t = k\} R_t, \quad \& \quad S_t^k := \sum_{s<t} \mathbb{1}\{A_t = k\} S_t.$$

Similarly, N_t^*, R_t^*, S_t^* denote the corresponding variables for k^*. Notice that $\mathcal{R}_t = \sum_{k \neq k^*} (\Delta^k \vee \Gamma^k) N_t^k$. We also use the notation $\widehat{\mu}_t^k := R_t^k / N_t^k, \widehat{\nu}_t^k := S_t^k / N_t^k$.

Since controlling it is of natural interest, we define the number of times an unsafe arm is played as

$$\mathcal{U}_T := \sum_t \mathbb{1}\{\nu^{A_t} > \alpha\}.$$

Finally, for $a, b \in [0, 1]$, we use the notation

$$d(a\|b) := a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$$

to denote the KL divergence between Bernoulli laws with means a and b. We will also need the notation

$$d_<(a\|b) := d(a\|b)\mathbb{1}\{a < b\},$$
$$d_>(a\|b) := d(a\|b)\mathbb{1}\{a > b\}.$$

Remark While the formulation focuses on a single safety-constraint, this may be extended. For example, we may posit a safety-risk vector $S \in [0,1]^d$, and demand that the corresponding (vector) means ν^k should lie in some known safe set $\mathcal{S}$. Natural extensions of the methods below would control, e.g., $\sum \max(\mu^{A_t} - \mu^*, \text{dist}(\nu^{A_t}, \mathcal{S}))$. We focus on a single constraint for clarity and ease of exposition.

3 Doubly Optimistic Confidence Bounds

The use of optimistic confidence bounds is well established in standard bandits (e.g. Ch. 7-10 Lattimore & Szepesvári, 2020). The idea is that pulling according to the maximum optimistic bound on the means encourages exploration, while efficiency follows because the confidence bounds exploit information to shrink towards the means, eventually giving evidence for the inefficiency of suboptimal arms.

The idea behind doubly optimistic bounds is identical - we maintain lower bounds on safety-risk L_t^k and upper bounds on rewards U_t^k such that $L_t^k \leq \nu^k$ and $U_t^k \geq \mu^k$ with high probability. We then construct a set of 'permissible arms' $\Pi_t := \{k : L_t^k \leq \alpha\}$ - these are all the arms that are plausibly feasible given the information we have up to time t. A_t is selected to maximise U_t^k amongst $k \in \Pi_t$. The optimism of Π_t allows us to explore for high rewards, but the concentration of L_t^k as N_t^k grows serves to identify unsafe arms, which then cease to be pulled. The broad scheme is described in Algorithm 1.

Algorithm 1 Doubly Optimistic Confidence Bounds

Input: K, functions U, L.
Initialise: $\mathscr{H}_0 \leftarrow \varnothing$
for $t = 1, 2, \ldots$ **do**
  if $t \leq K$ **then**
    $A_t \leftarrow t$
  else
    $\forall k, L_t^k \leftarrow L(t, \mathscr{H}_{t-1}, k)$.
    $\Pi_t \leftarrow \{k : L_t^k \leq \alpha\}$.
    $\forall k \in \Pi_t, U_t^k \leftarrow U(t, \mathscr{H}_{t-1}, k)$.
    $A_t \leftarrow \arg\max_{k \in \Pi_t} U_t^k$.
  end if
  Pull A_t, receive $(R_t, S_t) \sim \mathbb{P}^{A_t}$.
  Update $\mathscr{H}_t \leftarrow \mathscr{H}_{t-1} \cup \{(A_t, R_t, S_t)\}$.
end for

This scheme can be analysed using a variation of the standard bandit analysis. To control the play of unsafe arms, we argue that $\nu^k - L_t^k$ is bounded

as $\sqrt{\log(T)/N_t^k}$. Thus, if $\nu^k > \alpha$, the arm k should fall out of Π_t after it has been played at most $O(\log(T)/(\Gamma^k)^2)$ times. Next we argue that the bounds are 'consistent' (or optimistic) with high probabiliy, that is, most of the time $L_t^* \le \nu^* \iff k^* \in \Pi_t$ and $U_t^* \ge \mu^*$. Given this, in order to play arm k, U_t^k must exceed μ^*, but $U_t^k - \mu^k$ shrinks as $O(\sqrt{\log T/N_t^k})$ bounding N_T^k as $O(\log(T)/(\Delta^k)^2)$. In the process, strictly dominated k - for which $\nu^k > \alpha$ and $\mu^k < \mu^*$, are doubly penalised, and their play is limited by the larger gap.

We will explicitly analyse the scheme by instantiating the method with bounds based on KL-UCB (Garivier & Cappé, 2011), which offer optimal mean-dependent regret control for standard bandits. Note that the study of confidence bounds for bandit methods is mature, and our results can be improved with other choices of such bounds, e.g. , using variance sensitive bounds such as Empirical-KL-UCB (Cappé et al., 2013) or UCB_V (Audibert et al., 2009).

The KL-UCB type bounds take the following form

$$\begin{aligned}\gamma_t &:= \log(t(\log(t))^3),\\ U(t,\mathscr{H}_{t-1},k) &:= \max\{q > \widehat{\mu}_t^k : d(\widehat{\mu}_t^k \| q) \le \gamma_t/N_t^k\},\\ L(t,\mathscr{H}_{t-1},k) &:= \min\{q < \widehat{\nu}_t^k : d(\widehat{\nu}_t^k \| q) \le \gamma_t/N_t^k\},\end{aligned}$$

where γ_t trades-off the width and consistency of U, L. These bounds are natural for Bernoulli random variables, and since these are the 'least-concentrated' law on $[0,1]$, the fluctuation bounds extend to general random variables. Using these, we show the following result in §B.1.

Theorem 1. *Algorithm 1 instantiated with* KL-UCB *type bounds attains the following for any T and any $\varepsilon > 0$.*

$$\mathbb{E}[\mathcal{R}_T] \le \sum_{k \ne k^*} \frac{(1+\varepsilon)(\Delta^k \vee \Gamma^k)\log T}{d_<(\mu^k\|\mu^*) \vee d_>(\nu^k\|\alpha)} + \xi_k,$$

where $\xi_k = O(\log\log T + \varepsilon^{-2})$. Further, the number of times an unsafe arm is played is bounded as

$$\mathbb{E}[\mathcal{U}_T] \le \sum_{k:\Gamma^k>0} \left(\frac{(1+\varepsilon)\log T}{d_<(\mu^k\|\mu^*) \vee d_>(\nu^k\|\alpha)}\right) + \xi_k.$$

The O in the above hides instance-dependent constants, the most pertinent of which is a dependence on $(\Delta^k \vee \Gamma^k)^{-3}$ with the ε^{-2} term. To ameliorate this, we also give a gap-independent analysis of the scheme in §B.2.

Theorem 2. *Algorithm 1 instantiated with* KL-UCB *attains*

$$\mathbb{E}[\mathcal{R}_T] \le \sqrt{28KT\log T} + 6K\log\log T + 32.$$

The above statement extends to KL-UCB for standard bandits upon sending $\alpha \to 1$, which, surprisingly, appears to have been unobserved, at least explicitly.

4 Bayesian Methods

Thompson Sampling (TS) is the first proposed method for bandit problems (Thompson, 1933), and encourages exploration by using randomisation. The idea is to choose an benign prior, and play arms according to their posterior probability being optimal. The posteriors remain flat for insufficiently explored arms, giving a non-trivial chance of pulling them. An advantage of TS lies in the fact that it exploits a posterior that may be much better adapted to the underlying law $\mathbb{P}^k$ than confidence bounds that rely on a few simple

statistics. Indeed, it has been empirically observed that TS offers improved regret versus comparable UCB methods in multi-armed bandits (Chapelle & Li, 2011).

This section explores the use of Bayesian methods for safe bandits. We start by replacing the KL-UCB based selection of arms to play in Algorithm 1, but retaining the construction of Π_t. We then study a Bayesian method of selecting Π_t.

In the subsequent, we restrict analysis to the case of Bernoulli bandits, i.e., where the laws $\mathbb{P}^k$ are such that marginally $R \sim \mathrm{Bern}(\mu^{A_t})$ and $S \sim \mathrm{Bern}(\nu^{A_t})$. We note that since the resulting bounds depend on only the means of the rewards and safety-risk, these bounds extend to generic laws supported on $[0,1]^2$ - indeed, as observed by Agrawal & Goyal (2012), one can exploit an algorithm for Bernoulli bandits for generic laws by passing to the algorithm two samples $\widetilde{R}_t \sim \mathrm{Bern}(R_t), \widetilde{S}_t \sim \mathrm{Bern}(S_t)$. The corresponding $\widetilde{R}, \widetilde{S}$ are then Bernoulli with the same means, and any guarantee that only depends on the means for the Bernoulli case extends to the underlying bandit problem. Of course, such a procedure may blow up variances, and thus be profligate in the case of highly concentrated instances.

Note: the methods described below admit essentially the same guarantees as the bounds of Theorems 1 and 2. For the sake of brevity, we suppress the explicit bounds on $\mathbb{E}[\mathcal{U}_T]$ and the gap-independent bounds in the following.

4.1 Thompson Sampling with Optimistic Safety Indices

For Bernoulli bandits, it is natural to use the Beta family for priors, due to favourable conjugacy. The standard form of TS instantiates each arm with the uninformative prior $\mathrm{Beta}(1,1) = \mathrm{Unif}[0,1]$. The corresponding posterior at time t is $\mathrm{Beta}(R_t^k + 1, N_t^k - R_t^k + 1)$.

Algorithm 2 describes the proposed strategy - we retain the optimistic lower bound from Algorithm 1, but replace the arm selection given Π_t to a TS strategy: random scores ρ_t^k are drawn from the posterior for each arm in Π_t, the arm with the largest ρ_t^k is pulled.

Algorithm 2 Thompson Sampling With Optimistic Safety Indices (TOPSI) for Bernoulli Bandits

Input: K, function L.
Initialise: $\mathscr{H}_0 \leftarrow \varnothing$.
for $t = 1, 2, \ldots$ **do**
  if $t \leq N$ **then**
    $A_t \leftarrow t$
  else
    $\forall k, L_t^k \leftarrow L(t, \mathscr{H}_{t-1}, k)$.
    $\Pi_t \leftarrow \{k : L_t^k \leq \alpha\}$.
    $\forall k \in \Pi_t$, sample $\rho_t^k \sim \mathrm{Beta}(R_t^k + 1, N_t^k - R_t^k + 1)$
    $A_t \leftarrow \arg\max_{k \in \Pi_t} \rho_t^k$.
  end if
  Pull A_t, receive $(R_t, S_t) \sim \mathbb{P}^{A_t}$.
  Update $\mathscr{H}_t \leftarrow \mathscr{H}_{t-1} \cup \{(A_t, R_t, S_t)\}$.
end for

The analysis of such a method is simple, *given an analysis of* TS for standard bandits. Indeed, we can control the play of infeasible arms as we did for Algorithm 1. Further, as long as we can ensure $k^* \in \Pi_t$ with high probability, we can invoke the decomposition

$$\mathbb{E}[N_T^k] \leq \sum_t \mathbb{P}(k^* \notin \Pi_t) + \mathbb{P}(A_t = k, k^* \in \Pi_t).$$

The first term is handled using the consistency of the lower bound L_t^*. The second term is essentially the term analysed for standard bandits, and we can use any analysis of TS to control it. We concretely use the approach of Agrawal & Goyal (2013) in §C to show the following result.

Theorem 3. *For Bernoulli Bandits, Algorithm 2 instantiated with a* KL-UCB *type confidence bound attains the following regret bound for any T and any $\varepsilon > 0$*

$$\mathbb{E}[\mathcal{R}_T] \leq \sum_{k \neq k^*} \frac{(1+\varepsilon)(\Delta^k \vee \Gamma^k) \log T}{d_<(\mu^k \| \mu^*) \vee d_>(\nu^k \| \alpha)} + \xi_k,$$

where $\xi_k = O(\log\log T + \varepsilon^{-2} \log(1/\varepsilon))$

4.2 Thompson Sampling with BayesUCB

While Algorithm 2 admits a tight analysis, it still uses the potentially loose frequentist bound to decide Π_t, and it is possible that using the posteriors on the safety-risks to do this may improve the behaviour.

It is tempting to appeal to the basic structure of Thompson sampling, and associate a posterior with the safety risk of $P^k_{t,\nu} = \text{Beta}(S^k_t + 1, N^k_t - S^k_t + 1)$, sample safety scores $\sigma^k_t \sim P^k_{t,\nu}$, and let $\Pi_t = \{k : \sigma^k_t \leq \alpha\}$. However, this attempt is misguided, essentially because we need to compare the scores to a fixed level α, rather than amongst each other. Indeed, if it is the case that $\nu^* = \alpha$, then there is a constant chance that $\sigma^*_t > \alpha$, even if the empirical mean $\widehat{\nu}^*_t$ is faithful. This would mean a constant chance of playing a suboptimal arm, and so linear regret. A similar issue has been observed with trying to analyse TS using the analysis developed for UCB-type schema (Kaufmann et al., 2012b), but the issue is now at the level of the scheme rather than an analysis. Indeed, we show via simulations that when $\nu^* = \alpha$, such a scheme suffers linear expected regret (§F.4).

So, this idea needs a fix. One natural attempt is to introduce a slack, say some β^k_t, such that $\Pi_t = \{\sigma^k_t \leq \alpha + \beta^k_t\}$. This β^k_t should likely decay as N^k_t rises, but be large enough to ensure that $k^* \in \Pi_t$ - this is similar to the analytical approach taken by Kaufmann et al. (2012b). However, in designing such a β^k_t, we are functionally designing a confidence bound, somewhat defeating the purpose.

We take a different tack, and instead use a *Bayesian confidence bound*, essentially exploiting the BayesUCB method of Kaufmann et al. (2012a). The idea is to choose a δ^k_tth quantile of the posterior $P^k_{t,\nu}$ as a score, where δ^k_t is a schedule that decays with t. This is able to exploit the potentially improved adaptivity of the posterior, but due to δ^k_t being small, would continue to produce an optimistic score, and so have a high chance of $k^* \in \Pi_t$ at any time. Additionally, due to the concentration of the Beta-law for large N^k_t, the score of unsafe arms would converge towards ν^k, and thus preclude their play beyond a point. Altogether, the method seems tailor-made for our situation of filtering arms at a given level. The scheme is described in Algorithm 3, where $Q(P, \delta)$ denotes the δth quantile of the law P. We introduce a slight bias in the same for technical convenience.

Algorithm 3 Thompson Sampling with BayesUCB (tsbu) for Bernoulli Bandits

Input: K, schedule δ^k_t.
Initialise: $\mathscr{H}_0 \leftarrow \varnothing$.
for $t = 1, 2, \ldots$ **do**
  $\forall k$
  if $S^k_t = 0$ **then**
    $L^k_t \leftarrow 0$
  else
    $L^k_t \leftarrow Q(\text{Beta}(S^k_t, N^k_t - S^k_t + 1), \delta^k_t)$.
  end if
  $\Pi_t \leftarrow \{k : L^k_t \leq \alpha\}$.
  $\forall k \in \Pi_t$, sample $\rho^k_t \sim \text{Beta}(R^k_t + 1, N^k_t - R^k_t + 1)$
  $A_t \leftarrow \arg\max_{k \in \Pi_t} \rho^k_t$.
  Pull A_t, receive $(R_t, S_t) \sim \mathbb{P}^{A_t}$.
  Update $\mathscr{H}_t \leftarrow \mathscr{H}_{t-1} \cup \{(A_t, R_t, S_t)\}$.
end for

The main design parameter is δ^k_t which trades off the consistency and tightness. In our argument, we use a conservative choice of $\delta^k_t = (\sqrt{8N^k_t} t \log^3 t)^{-1}$, which leads to a simplified proof, but introduces the inefficiency of $2/3$ in the bounds below. We find that in simulations, the uniform choice $\delta^k_t = 1/(t+1)$ is better (§F.3), and perhaps an improved analysis can establish better bounds for such a schedule. The following summarises our analysis in §D.

Theorem 4. *For Bernoulli bandits, Algorithm 3, instantiated with* $\delta^k_t = (\sqrt{8N^k_t} t \log^3 t)^{-1}$ *attains the following regret bound for any* $\varepsilon > 0$ *and any* T :

$$\mathbb{E}[\mathcal{R}_T] \leq \sum_{k \neq k^*} \frac{(1+\varepsilon)(\Delta^k \vee \Gamma^k) \log T}{d_<(\mu^k \| \mu^*) \vee 2/3 \cdot d_>(\nu^k \| \alpha)} + \xi_k,$$

where $\xi_k = O(\log \log T + \varepsilon^{-2} \log(1/\varepsilon))$

5 Lower Bound

We conclude our theoretical study with a lower bound for algorithms that admit sub-polynomial regrets against all bounded distributions. This is based on the technique of Garivier et al. (2019), who use the chain rule of KL divergence and the data processing inequality to show the following relation, which extends to our case without change:

Lemma 5. *For any safe bandit algorithm, and any two safe bandit instances* $\{\mathbb{P}^k\}, \{\widetilde{\mathbb{P}}^k\}$, *and any* T,

$$\sum \mathbb{E}[N_T^k] D(\mathbb{P}^k \| \widetilde{\mathbb{P}}^k) \geq d(\mathbb{E}[N_T^k/T] \| \widetilde{\mathbb{E}}[N_T^k/T]).$$

This lemma enables a standard approach - pick $\widetilde{\mathbb{P}}$ so that $\widetilde{\mathbb{E}}^k[(R,S)] = (\mu^k \vee \mu^* + \varepsilon, \nu^k \wedge \alpha)$, and leave the other $\mathbb{P}^k$s unchanged. For any bandit algorithm with sub-polynomial mean regret, the right hand side grows as $\log(T)$, while the left hand side side reduces to $\mathbb{E}[N_T^k] D(\mathbb{P}^k \| \widetilde{\mathbb{P}}^k)$. Of course, the optimal choice of $\widetilde{\mathbb{P}}$ depends subtly on the details of $\mathbb{P}$. We study a simple concrete case to illustrate that our prior analyses are tight.

Proposition 6. *Any algorithm that ensures that, uniformly over all instances of safe Bernoulli bandit problems with independent rewards and safety-risks, the mean number of plays of any suboptimal arm is bounded as* $O(T^x)$ *for every* $x \in (0,1)$ *must satisfy*

$$\varliminf_{T \nearrow \infty} \frac{\mathbb{E}[N_T^k]}{\log T} \geq \frac{1}{d_<(\mu^k \| \mu^*) + d_>(\nu^k \| \alpha)}.$$

Since mean regret can be expressed in terms of $\mathbb{E}[N_T^k]$, this also lower bounds regret. Note the sum in the denominator, rather than a max as in our upper bounds. This means that for strictly dominated arms (i.e. $k : \Delta^k > 0, \Gamma^k > 0$), our scheme may be loose by up to a factor of two. This arises since our scheme does not utilise the dependence structure of (R,S), and represents an opportunity for future work.

6 Simulations

We provide practical contextualisation for the schema described in the theoretical section using simulation studies over small safe bandit environments. Of course, to concretely study the schema we describe, we need to instantiate them with appropriate confidence bounds. We will do so using the KL-UCB and BayesUCB based indices which we analysed previously. Finally, we use TS instantiated with the Beta priors as described in the text. Of course, a variety of other methods can be implemented in these schema, but we believe that these methods serve well to illustrate both the theory and a first order practical design. All implementation details are left to §F.

We will begin by empirically illustrating that the prior policy based methods are indeed ineffective in our scenario, and play unsafe arms far too often. We then illustrate the performance of the methods on a realistic problem instance. Finally, we will and investigate the dependence of regret of the three methods on the gaps to the optimal arm.

6.1 Empirical Demonstration of the Ineffectiveness of Prior Formulations

As discussed previously, the globally constrained (Badanidiyuru et al., 2013; Agrawal & Devanur, 2014) and policy level (Pacchiano et al., 2021) formulations are unsatisfactory in the context of safety-constraints, as illustrated by example (2). Nevertheless, a priori it may be possible that the schema designed for these objectives may be effective in our scenario, especially if there exist optimal policies supported on a single arm. We implement the the doubly optimistic policy method (BwCR) of Agrawal & Devanur (2014), and the optimistic-pessimistic method (Pess) of Pacchiano et al. (2021) to demonstrate that this is untrue.

We explore two illustrative cases, both of which are for Bernoulli bandits with independent means and safety-risks. The data reported is across 100 trials of horizon 50000. Since these policy methods are based on confidence bounds, we also compare them to Alg. 1. In all cases we instantiate these schema with KL-UCB-based confidence bounds.

1. Multiple optimal policies. We consider four arms with

$$\mu = (0, 0.4, 0.5, 0.6), \nu = (0, 0.4, 0.5, 0.6), \alpha = 0.5.$$

The $(0, 0)$ arm is included as a known safe arm, which is required for Pess to enable sufficient exploration. Notice that in this case there are two optimal static policies - one that is entirely supported on arm 3, while another that is uniformly supported on arms 2 and 4. However, which one of these two policies these schema converge to is essentially random, and we thus see linear growth of $\mathbb{E}[\mathcal{U}_t]$ in Fig. 1.

2. Single arm optimal policy. We jack up the rewards of arm 3 to 0.6, but leave the other means unchanged. Now the optimal policy is singly supported on arm 3 and has a significant gap of 0.1. Despite the fact that such a case is the most promising for policy-based methods in terms of efficacy in our formulation, Fig. 1 again shows that they do rather poorly - for instance, while our implementation play the unsafe arms about 550 times, these methods play it at least 8000 times. This occurs because the policy-based methods are designed for the much richer policy space—a simplex—and so must explore a lot more than methods designed for single arm play. We note that in this case, while BwCR plays unsafe arms more often, it suffers less regret than Pess, since the unsafe arm incurs a smaller loss.

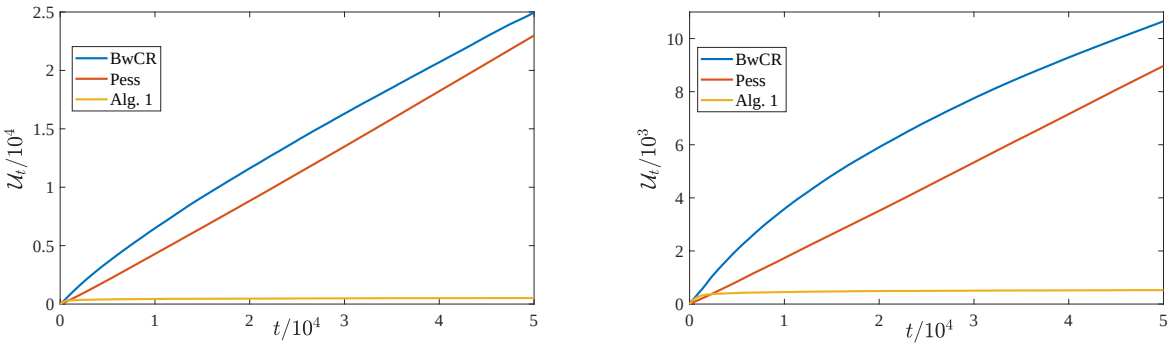

Figure 1: Empirical means of $\mathcal{U}_t$ versus t averaged over 100 trials over $t \in [1 : 50000]$. Left is case 1 (with multiple optimal policies), right case 2 (with a single optimal policy supported on one arm).

6.2 Characterisation of the Proposed Schema

We implement the three methods to establish a practical contextualisation of their performance, and to verify the theoretical claims. For the sake of realism, we use the data of Genovese et al. (2013), who report efficacy and infection rates from a phase 2 randomised trial for various dosages of a drug to treat rheumatoid arthritis. The dosages studied were $(0, 25, 75, 150, 300)$ mg, and the observations were

$$\mu = (0.360, 0.340, 0.469, 0.465, 0.537),$$
$$\nu = (0.160, 0.259, 0.184, 0.209, 0.293).$$

This data is challenging for any safety level - no matter the choice, we have to deal with either a potential safety gap of order 10^{-2}, or an efficacy gap of 10^{-3}, both of which contribute a large regret. We study the safety level 0.21, under which arm 3 is optimal, while arms $2, 5$ are unsafe. We chose this to allow large enough safety gaps that the behaviour of $\mathcal{U}_T$ is easy to establish with runs of length about $50K$ - if we took α smaller, say 0.2, then we would expect to need runs of length $100K$ simply to reach a point at which arm 4 is played fewer than about a third of the time. This consideration also illustrates why the regret $\mathcal{R}_T$ is a much

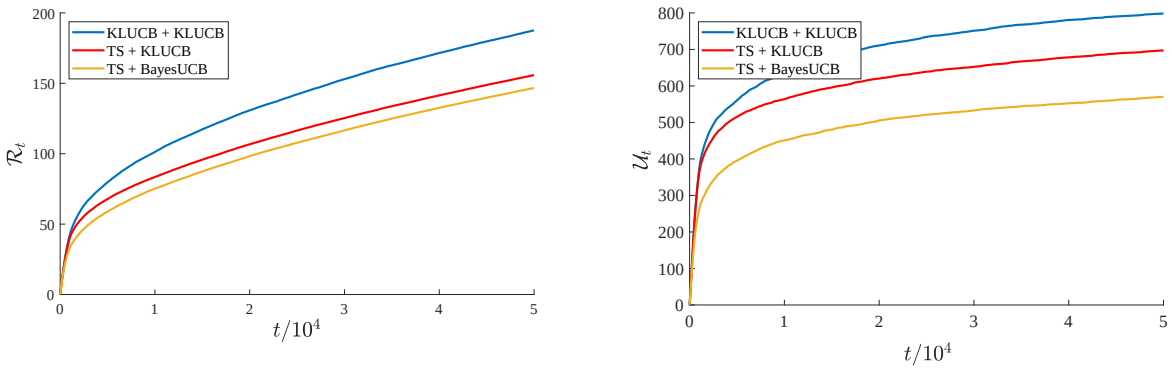

Figure 2: Empirical means over 500 trials of $\mathcal{R}_t$ (left) and $\mathcal{U}_t$ (right) for the drug trial data with $\alpha = 0.21$.

more reasonable notion of study than $\mathcal{U}_T$, which can grow very large due to tiny, practically undetectable safety gaps. Plots for a run with $\alpha = 0.19$ are included in §F.2.

Observations of Performance From Fig.2, we first note that both $\mathcal{R}_t$ and $\mathcal{U}_t$ are well controlled and well within the theoretical bounds for the methods we have analysed.[1] The general trend observed is that Alg. 2 based methods that use a TS-based index outperform confidence bound indices of Alg. 1, which is consistent with Chapelle & Li (2011). Finally, we observe that Alg. 3, as represented by TS+BayesUCB outperforms all other methods. These observations held regardless of the means we have run the methods on. One caveat, however, is that the underlying Bernoulli laws used are well aligned to the priors for Bayesian methods, which may improve their performance.

6.2.1 Dependence on Gaps

We investigate the dependence of the regret on the gaps Δ^k, Γ^k, in particular illustrating that, as predicted by the theorems, this decays inversely with the larger of the two, and is insensitive to the smaller of the two.

Inverse Dependence on Gaps First, we will demonstrate that the regret varies with $(\Delta^k \wedge \Gamma^k)$ inversely. To this end, we study the the cases

$$\mu_i = (0.5, 0.5 - i/25, 0.5 + i/25),$$
$$\nu_i = (0.5, 0.5 - i/25, 0.5 + i/25),$$

for $\alpha = 0.5$ over i in $[1:10]$ over 100 trials across a horizon of $T = 2 \times 10^4$. Fig. 3 reports the regret $\mathcal{R}_T$ versus $i/25$ over this data, and exhibits a clear inverse dependence on i.

Lack of Dependence on Smaller Gaps Secondly, we will illustrate that the dependence on the gaps is driven by the *larger* of Δ^k and Γ^k, but not on $(\Delta^k \wedge \Gamma^k)$. For this we study the data

$$\mu_i = (0.5, 0.5 - i/25, 0.5 + i/250),$$
$$\nu_i = (0.5, 0.5 + i/250, 0.5 + i/25),$$

again with $\alpha = 0.5$ for 100 trials over a horizon of $T = 2 \times 10^4$. Observe that $\Delta^k \vee \Gamma^k$ is the same as the previous case, but $\Delta^k \wedge \Gamma^k$ is reduced by a factor of 10 for each suboptimal arm. The principal observation from the second part of Fig. 3 is that the plot remains similar to the previous case of 'large' minimum gaps, bearing out this independence from the smaller of the two gaps.

[1] The main term of the regret bound is $137 \log t$, and the unsafe-arm bound is $81 \log t$, both > 750 for $t > 10^4$.

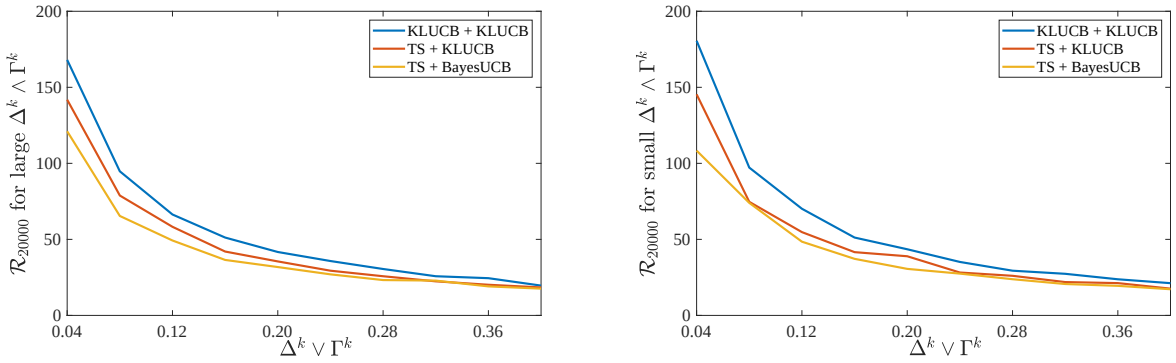

Figure 3: Behaviour of Regret at $T = 20000$ with respect to the maximum gap. Medians over 100 runs are reported. Left is the case of large minimum gaps, while right is the case of small minimum gaps.

References

Agrawal, S. and Devanur, N. R. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014. 3, 9

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012. 7

Agrawal, S. and Goyal, N. Further optimal regret bounds for Thompson sampling. In *Artificial intelligence and statistics*, pp. 99–107. PMLR, 2013. 7, 15, 20, 21, 22

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 3

Audibert, J.-Y., Munos, R., and Szepesvári, C. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009. 6

Aziz, M., Kaufmann, E., and Riviere, M.-K. On multi-armed bandit designs for dose-finding clinical trials. *Journal of Machine Learning Research*, 22:1–38, 2021. 4

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 207–216. IEEE, 2013. 3, 9

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pp. 1516–1541, 2013. 6, 19

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011. 7, 11

Garivier, A. and Cappé, O. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011. 6, 15, 19, 27, 30

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019. 9

Genovese, M. C., Durez, P., Richards, H. B., Supronik, J., Dokoupilova, E., Mazurov, V., Aelion, J. A., Lee, S.-H., Codding, C. E., Kellner, H., et al. Efficacy and safety of secukinumab in patients with rheumatoid arthritis: a phase ii, dose-finding, double-blind, randomised, placebo controlled study. *Annals of the rheumatic diseases*, 72(6):863–869, 2013. 1, 10

Jeřábek, E. Dual weak pigeonhole principle, boolean complexity, and derandomization. *Annals of Pure and Applied Logic*, 129(1-3):1–37, 2004. 26

Katz-Samuels, J. and Scott, C. Feasible arm identification. In *International Conference on Machine Learning*, pp. 2535–2543. PMLR, 2018. 4

Katz-Samuels, J. and Scott, C. Top feasible arm identification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1593–1601. PMLR, 2019. 4

Kaufmann, E., Cappé, O., and Garivier, A. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pp. 592–600. PMLR, 2012a. 8, 15, 24, 28, 30

Kaufmann, E., Korda, N., and Munos, R. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer, 2012b. 8, 20

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020. 2, 5

Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. Safe linear Thompson sampling with side information. *IEEE Transactions on Signal Processing*, 2021. 3

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 2827–2835. PMLR, 2021. 3, 9

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. 6

Wang, Z., Wagenmaker, A., and Jamieson, K. Best arm identification with safety constraints. *arXiv preprint arXiv:2111.12151*, 2021. 4

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. Conservative bandits. In *International Conference on Machine Learning*, pp. 1254–1262. PMLR, 2016. 3

A Notation and Broad Proof Strategy

We begin with some notation, and then describe the general proof strategy.

We will abuse notation and let $\mathscr{H}_{t-1}$ also stand for the sigma algebra induced by the history, with $\mathscr{H}_0$ denoting the trivial sigma algebra. Naturally, $\{\mathscr{H}_t\}$ forms a filtration - observe that in the TS cases, the laws of ρ_t^k are measurable with respect to $\mathscr{H}_{t-1}$. The Bayesian methods also utilise extraneous randomness, as represented by the various ρ_t^ks. An important observation regarding all the methods we design is that the permissible set Π_t is a *predictable* process, i.e., is determined given $\mathscr{H}_{t-1}$. Indeed, all the methods use an index based on the history to decide Π_t, and so it is a deterministic function of the variables $\{(A_s, R_s, S_s) : s < t\}$. Of course, this is not as such required, but it represents a convenience that we will employ in our proofs. For the sake of brevity, we will denote the conditional laws $\mathbb{P}(\cdot|\mathscr{H}_{t-1})$ as $\mathbb{P}_{t-1}$.

Proof strategy The basic decomposition of regret is in terms of N_T^k - indeed, due to the additive definition,

$$\mathbb{E}[\mathcal{R}_T] = \sum_{k \neq k^*} \mathbb{E}[N_T^k](\Delta^k \vee \Gamma^k).$$

Therefore, the main arguments all control $\mathbb{E}[N_T^k]$ for all suboptimal arms k. Of course, subsidiary claims about $\mathbb{E}[\mathcal{U}_T]$ also follow from these.

The arguments separately control $\mathbb{E}[N_T^k]$ for infeasible and inefficient arms. For arms which are both inefficient and infeasible, the tighter of the control offered by these two arguments can be taken, and this yields the form of the expressions in the main text.

Infeasible arms All of our schemes use a safety index L_t^k to populate the permissible set Π_t. We exploit the properties of this index to control the play of infeasible arms. Indeed, we can decompose

$$\mathbb{E}[N_T^k] = \sum \mathbb{P}(A_t = k) \leq \sum_{t=1}^{T} \mathbb{P}(L_t^k \leq \alpha).$$

The design of the two indices - that via KL-UCB and via BayesUCB both ensure that the chance of playing an infeasible arm more than $O(\log(T)/d(\nu^k\|\alpha))$ times is exponentially small. For KL-UCB, this is a simple consequence of Chernoff's bound. For BayesUCB, the argument reduces to that for KL-UCB using a connection between the tails of Beta distributions and Binomials.

Inefficient arms Following the standard method for confidence bound based index policies, controlling the play of inefficient arms requires some known good index to compare the reward indices to. Naturally, we want to use the index of k^*, but doing so requires that k^* itself is permitted, since otherwise the algorithm never takes its reward index into consideration when choosing an arm. This represents the main deviation from standard proofs.

Let us take the case of KL-UCB. The idea is to decompose

$$\begin{aligned}
\mathbb{E}[N_t^K] &= \sum \mathbb{P}(A_t = k) \\
&= \sum \mathbb{P}(A_t = k, k^* \notin \Pi_t) + \mathbb{P}(A_t = k, k^* \in \Pi_t) \\
&\leq \sum_t \mathbb{P}(k^* \notin \Pi_t) + \sum_t \mathbb{P}(A_t = k, k^* \in \Pi_t)
\end{aligned}$$

The first course of action then is to ensure that the first term is small, which exploits the consistency of L_t^*.

This enables us to proceed pretty much as usual. For KL-UCB, we decompose the second term as

$$\begin{aligned}\sum_t \mathbb{P}(A_t = k, k^* \in \Pi_t) &\leq \sum_t \mathbb{P}(U_t^* \leq \mu^*, k^* \in \Pi_t) + \mathbb{P}(U_t^k < \mu^*, U_t^* \geq \mu^*, k^* \in \Pi_t, A_t = k) \\ &\leq \sum_t \mathbb{P}(U_t^* < \mu^*) + \mathbb{P}(U_t^k \geq \mu^*, A_t = k).\end{aligned}$$

Of course, the final expression is the usual quantity controlled in regret proofs, and this argument can be repeated without change. For the sake of being self-contained, we will sketch these proofs in the subsequent as well. For KL-UCB, this is essentially the argument of Garivier & Cappé (2011), while for the BayesUCB bound, this is the argument of Kaufmann et al. (2012a) (which itself is very similar to Garivier & Cappé (2011)). For the efficiency of TS, we will use the argument of Agrawal & Goyal (2013).

Remark on showing consistency of L_t^* We observe that, by design, our choices of L_t^k are such that consistency proofs for U_t^* translate directly into those for L_t^* - this is due to the symmetry of the relevant functionals under the maps $(S, \nu^k, \alpha) \mapsto (1 - S, 1 - \nu^k, 1 - \alpha)$, upon doing which $1 - L_t^k$ is a U_t^k-type upper bound for $1 - \nu^k$. Similarly, the argument for controlling $\sum \mathbb{P}(L_t^k \leq \alpha)$ for infeasible arms is basically the same as that for controlling $\sum \mathbb{P}(U_t^k \geq \mu^*)$ for the standard bandit version of the appropriate method.

That said, we note a deviation from the proof of this consistency for the case of BayesUCB. Since controlling standard regret in a Bayesian setting requires one to compare two *random* indices, Kaufmann et al. (2012a) use direct comparison of their index $U^*_{t,\text{BayesUCB}}$ to μ^* only enough to argue that N_t^* is at least logarithmically large. With this in hand, they can argue that $U^*_{t,\text{BayesUCB}}$ is at least $\mu^* - O(\sqrt{1/\log(T)})$ with high probability and argue that this is unlikely to be exceeded by suboptimal arms. However, to ensure that our (random) safety index L_t^* is consistent, we must compare it to a fixed value α, and so this second argument utilising a weakened consistency does not carry over. We handle this by loosening the quantiles δ_t^k enough so that the first argument itself is sufficient to provide consistency. This represents a gap, which may be possibly be resolved by a stronger analysis.

Remark on dependence Note that the sketch above above does not use the potential dependence between the signals (R, S). It is possible that this can be exploited, and this exploitation may gain in importance as we increase the number of safety constraints. We leave this as a direction for further work.

B Proof for Doubly Optimistic Confidence Bounds

The following lemma essentially follows from the main result of the KL-UCB analysis due to Garivier & Cappé (2011), and forms the key statement to demonstrate our results. We note that this is stated slightly more generically than in their paper, essentially to let us use the same result to show both gap dependent and gap independent bounds. We came across this trick in the work of Agrawal & Goyal (2013).

Lemma 7 (Adaptation of Garivier & Cappé (2011))**.** *Let k be a suboptimal arm. Then Algorithm 1, instantiated with the* KL-UCB *type confidence bounds attains the following guarantees for all k.*

- *If $\Delta^k > 0$, then for any $x \in (\mu^k, \mu^*)$,*

$$\mathbb{E}[N_T^k] \leq \frac{\log T + 3\log\log T}{d(x\|\mu^*)} + 6\log\log T + \frac{2}{1 \wedge d(x\|\mu^k)} + 24. \tag{3}$$

- *If $\Gamma^k > 0$, then for any $y \in (\alpha, \nu^k)$,*

$$\mathbb{E}[N_T^k] \leq \frac{\log T + 3\log\log T}{d(y\|\alpha)} + \frac{2}{1 \wedge d(y\|\nu^k)}. \tag{4}$$

We will first show the proofs of the two results using the above lemma, and leave proving it until the end.

B.1 Proof of Theorem 1

Proof. Fix an arm k. If $\Delta^k > 0$, then choose $x \in (\mu^k, \mu^*)$ such that $d(x\|\mu^*) = \frac{d(\mu^k\|\mu^*)}{1+\varepsilon}$ - this exists since $d(x\|\mu^*)$ is continuous and monotonically decreases from $d(\mu^k\|\mu^*)$ to 0 as x varies in (μ^k, μ). We need to argue that the third term in the bound of (3) is bounded as $O(\varepsilon^{-2})$. This follows since for small ε, $x = \mu^k + O(\varepsilon)$.

Indeed, let us abbreviate $d = d(\mu^k\|\mu^*)$, and observe that the the derivative $d' := \partial_z d(z\|\mu^*)|_{z=\mu^k}$ is non-zero, and so $x - \mu^k = \varepsilon\frac{d}{|d'|} + O(\varepsilon^2)$. But then notice that since $d(z\|\mu^k)$ is minimised at $z = \mu^k$, $d(x\|\mu^k) = 1/2(\widetilde{d}''\varepsilon^2(d/d')^2) + O(\varepsilon^3)$, where $\widetilde{d}'' := \partial^2_{zz} d(z\|\mu^k)\big|_{z=\mu^k}$. We conclude that

$$\frac{2}{d(x\|\mu^k) \wedge 1} = O\left(\frac{d'^2}{d''d^2\varepsilon^2}\right),$$

which of course is a scaling of ε^{-2} by a problem dependent constant.

Next, if $\Gamma^k > 0$, we proceed similarly to the above, and choose $y \in (\alpha, \nu^k)$ such that $d(y\|\alpha) = d(\nu^k\|\alpha)/(1+\varepsilon)$. By an entirely identical calculation as above, the final term of (4) is bounded as $O\left(\frac{f'^2}{\widetilde{f}''}\frac{1}{d^2(\nu^k\|\alpha)\varepsilon^2}\right)$, where $f' = \partial_z d(z\|\alpha)|_{z=\nu^k}$, and $\widetilde{f}'' = \partial^2_{zz} d^2(z\|\nu^k)\big|_{z=\nu^k}$.

Using both of these bounds, we conclude that

$$\mathbb{E}[N_T^k] \le \frac{1}{\mathbb{1}\{\mu^k < \mu^*\}}\left\{\frac{(1+\varepsilon)\log T}{d(\mu^k\|\mu^*)} + \frac{(1+\varepsilon)3\log\log T}{d(\mu^k\|\mu^*)} + 6\log\log T + 24 + O\left(\frac{(d'^2/\widetilde{d}'')}{d^2(\mu^k\|\mu^*)\varepsilon^2}\right)\right\},$$

$$\mathbb{E}[N_T^k] \le \frac{1}{\mathbb{1}\{\nu^k > \alpha\}}\left\{\frac{(1+\varepsilon)\log T}{d(\nu^k\|\alpha)} + \frac{(1+\varepsilon)3\log\log T}{d(\nu^k\|\alpha)} + O\left(\frac{(f'^2/\widetilde{f}'')}{d^2(\nu^k\|\alpha)\varepsilon^2}\right)\right\},$$

where we set $1/\mathbb{1}\{p\} = \infty$ when the proposition p is untrue. Of course, recalling that $\mathbb{1}\{\mu^k < \mu^*\}d(\mu^k\|\mu^*) = d_<(\mu^k\|\mu^*)$ and similarly $d_>(\nu^k\|\nu^*)$, we may choose the tighter of the above bounds to get the result

$$\mathbb{E}[N_t^k] \le \frac{(1+\varepsilon)\log T}{d_<(\mu^k\|\mu^*) \vee d_>(\nu^k\|\nu^*)} + O\left(\frac{\log\log T}{d_<(\mu^k\|\mu^*) \vee d_>(\nu^k\|\nu^*)} + \frac{1}{(d_<(\mu^k\|\mu^*) \vee d_>(\nu^k\|\nu^*))^2\varepsilon^2}\right).$$

The claimed bounds now follow trivially - to control $\mathbb{E}[\mathcal{R}_T]$, simply multiply by the per-round regret of playing arm k, $\Delta^k \vee \Gamma^k$, and sum. To control $\mathbb{E}[\mathcal{U}_T]$, simply add up the above over the unsafe arms.

□

Note that as the gaps Δ^k and Γ^k decay, the last term scales as $1/(\Delta^k \wedge \Gamma^k)^4$, which only yields a $T^{3/4}$ gap-independent bound.

B.2 Proof of Theorem 2

As is standard, the gap-independent regret bounds follow on observing that arms for which the gap is too small cannot actually incur large regret over T rounds. To this end, let $\mathbf{M} > 0$ be a parameter to be chosen, and express regret as

$$\mathbb{E}[\mathcal{R}_T] \le \sum_{k:\Delta^k > \Gamma^k \vee \mathbf{M}} \mathbb{E}[N_T^k]\Delta^k + \sum_{k:\Gamma^k > \Delta^k \vee \mathbf{M}} \mathbb{E}[N_T^k]\Gamma^k + \mathbf{M}\sum_{k:(\Delta^k \vee \Gamma^k) \le \mathbf{M}} \mathbb{E}[N_T^k]. \tag{5}$$

The last term is of course bounded by $\mathbf{M}T$, and so we will end up taking $\mathbf{M}$ of order $\sqrt{K\log T/T}$ to control regret. It remains to show that $\mathbb{E}[N_T^k]$ is not too large for arms with large gaps. To this end, we first develop bounds dependent explicitly on the gaps using (3) and (4).

Lemma 8. *For any arm k with $\Delta^k > 0$,*

$$\mathbb{E}[N_T^k] \leq \frac{2\log T + 6\log\log T + 4}{(\Delta^k)^2} + 6\log\log T + 24.$$

Similarly, for any arm k with $\Gamma^k > 0$,

$$\mathbb{E}[N_T^k] \leq \frac{2\log T + 6\log\log T + 4}{(\Gamma^k)^2}.$$

Proof. First, take a k with $\Delta^k > 0$, and in the bound (3), set $x = (\mu^k + \mu^*)/2 =: \bar{\mu}^k$. By Pinsker's inequality, $d(\bar{\mu}^k \| \mu^*) \geq 2(\mu^* - \bar{\mu}^k)^2 = (\Delta^k)^2/2$, and $d(\bar{\mu}^k \| \mu^k) \geq 2(\bar{\mu}^k - \mu^k)^2 = (\Delta^k)^2/2$. Plugging these into the bound yields the claim upon observing that $(\Delta^k)^2/2 \leq 1$.

For arms with $\Gamma^k > 0$, we can develop a similar control resulting from (4) by setting $y = (\alpha + \nu^k)/2$. □

We are now in a position to show the claim.

Proof of Theorem 2. The first term in (5) can be bounded as

$$\sum_{\Delta^k > \Gamma^k \vee \mathbf{M}} \frac{2\log T + 6\log\log T + 2}{\Delta^k} + (6\log\log T + 24)\Delta^k \leq K_\Delta \left(\frac{2\log T + 6\log\log T + 4}{\mathbf{M}} + 6\log\log T + 24 \right),$$

where $K_\Delta = |\{k : \Delta^k > \Gamma^k\}|$.

Similarly, the second term in (5) can be bounded as

$$\sum_{\Gamma^k > \Delta^k \vee \mathbf{M}} \frac{2\log T + 6\log\log T + 4}{\Gamma^k} \leq K_\Gamma \frac{2\log T + 6\log\log T + 4}{\mathbf{M}},$$

where $K_\Gamma = |\{k : \Gamma^k > \Delta^k\}|$.

Finally, observing that $K_\Gamma + K_\Delta \leq K$, we conclude that

$$\mathbb{E}[\mathcal{R}_T] \leq \frac{K}{\mathbf{M}}(2\log T + 6\log\log T + 4) + (6\log\log T + 24)\sum(\Delta^k \vee \Gamma^k) + T\mathbf{M}.$$

The claim follows on choosing $\mathbf{M} = \sqrt{K(2\log T + 6\log\log T + 4)/T}$, and observing that $2\log T \geq 4$ for $T \geq 8$, and $2\log\log T \leq \log T$ for all T. □

B.3 Proof of Lemma 7

Proof. We make the argument separately for infeasible and inefficient arms. The former is easier, so let us begin with it.

Infeasible arms

We follow the decomposition from §A. Recall that $L_t^k = \min\{q \leq \widehat{\nu}_t^k : d(\widehat{\nu}_t^k \| q) \leq \gamma_t / N_t^k\}$. Since $d(\widehat{\nu}_t^k \| x)$ is a continuous decreasing function on $[0, \widehat{\nu}_t^k]$, if $L_t^k \leq \alpha$ then it must either hold that $\widehat{\nu}_t^k \leq \alpha$, or $d(\widehat{\nu}_t^k \| \alpha) \leq d(\widehat{\nu}_t^k \| L_t^k) = \gamma_t / N_t^k$. Either way, we have that $d_>(\widehat{\nu}_t^k \| \alpha) \leq \gamma_t / N_t^k$.

Now, let $\hat{\nu}^k(s)$ denote the value of $\widehat{\nu}_t^k$ after the sth time we play the arm k. We observe that

$$\begin{aligned}
\sum_t \mathbb{1}\{A_t = k\} &\leq \sum_{t=1}^{T} \mathbb{1}\{A_t = k, d_>(\widehat{\nu}_t^k \| \alpha) \leq \gamma_t / N_t^k\} \\
&= \sum_{t=1}^{T} \sum_{s=1}^{t} \mathbb{1}\{A_t = k, sd_>(\widehat{\nu}_t^k \| \alpha) \leq \gamma_t, N_t^k = s\} \\
&\leq \sum_{t=1}^{T} \sum_{s=1}^{t} \mathbb{1}\{A_t = k, N_t^k = s\} \cdot \mathbb{1}\{sd_>(\hat{\nu}^k(s) \| \alpha) \leq \gamma_T\} \\
&= \sum_{s=1}^{T} \mathbb{1}\{sd_>(\hat{\nu}^k(s) \| \alpha) \leq \gamma_T\} \cdot \sum_{t=s}^{T} \mathbb{1}\{A_t = k, N_t^k = s\} \\
&= \sum_{s=1}^{T} \mathbb{1}\{sd_>(\hat{\nu}^k(s) \| \alpha) \leq \gamma_T\},
\end{aligned}$$

where we have used that γ_t increases with T, and for any value s, there is at most one time step on which N_t^k is exactly s and we play the action k.

Now, we observe that for any $y \in (\alpha, \nu^k)$, the event $\{d_>(\hat{\nu}^k(s) \| \alpha) \leq d(y \| \alpha)\} = \{\hat{\nu}^k(s) \leq y\}$. Indeed, $d_>(u \| \alpha)$ is exactly equal to 0 for $u \leq \alpha$, and monotonically increasing for $u > \alpha$. But, recalling Chernoff's bound (which applies since the random variables are bounded in $[0, 1]$), $P(\hat{\nu}^k(s) \leq y) \leq \exp(-sd(y \| \nu^k))$. This sets up the following calculation.

Let $y \in (\alpha, \nu^k)$, and define $S(y) := \lfloor \gamma_T / d(y \| \alpha) \rfloor$, so that for all $s > S(y)$, $\gamma_T / s < d(y \| \alpha)$. Then

$$\begin{aligned}
\mathbb{E}[N_T^k] &= \sum_{t=1}^{T} \mathbb{P}(A_t = k) \\
&\leq \sum_{s=1}^{T} \mathbb{P}(sd_>(\hat{\nu}^k(s) \| \alpha) \leq \gamma_t) \\
&\leq S(y) + \sum_{s=S(y)+1}^{T} \mathbb{P}(d_>(\hat{\nu}^k(s) \| \alpha) \leq d(y \| \alpha)) \\
&\leq S(y) + \sum_{s=S(y)+1}^{T} e^{-sd(y \| \nu^k)} \\
&\leq S(y) + \frac{e^{-(S(y)+1)d(y \| \nu^k)}}{1 - e^{-d(y \| \nu^k)}} \\
&\leq S(y) + \frac{2}{1 \wedge d(y \| \nu^k)}, \qquad (6)
\end{aligned}$$

where the last term uses that $(S(y) + 1)d(y \| \nu^k) \geq 0$, and $\frac{1}{1-e^{-u}} \leq 2/(1 \wedge u)$. But $S(y) \leq \frac{\gamma_T}{d(y \| \alpha)} = \frac{\log T + 3 \log \log T}{d(y \| \alpha)}$.

Inefficient arms Again, we follow the decomposition from §A, namely

$$\mathbb{E}[N_T^k] \leq \sum_t \mathbb{P}(k^* \notin \Pi_t) + \mathbb{P}(U_t^* < \mu^*) + \mathbb{P}(U_t^k \geq \mu^*).$$

Observe that $\mathbb{P}(k^* \notin \Pi_t) = \mathbb{P}(L_t^* > \alpha)$.

As noted in §A, the final term is controlled in exactly the same way as the inefficiency control. Indeed, $U_t^k = \max\{q \geq \widehat{\mu}_t^k : d(\widehat{\mu}_t^k \| q) \leq \gamma_t / N_t^k\}$. Since $d(\widehat{\mu}_t^k \| x)$ increases in the range $[\widehat{\mu}_t^k, 1]$, if $U_t^k \geq \mu^*$, then either $\widehat{\mu}_t^k > \mu^*$, or $d(\widehat{\mu}_t^k \| \mu^*) \leq \gamma_t / N_t^k$. Developing the subsequent bound in exactly the same way, we find that

$$\sum_t \mathbb{1}\{A_t = k, U_t^k \geq \mu^*\} \leq \sum_{s=1}^{T} \mathbb{1}\{sd_<(\hat{\mu}^k(s) \| \mu^*) \leq \gamma_T\},$$

and again, for any $x \in (\mu^k, \mu^*)$, $P(d_<(\hat{\mu}^k(s) \| \mu^*) \leq d(x \| \mu^*)) = P(\hat{\mu}^k(s) \leq x) \leq \exp(-d(x \| \mu^k))$. The resulting sum then gives the bound

$$\sum \mathbb{P}(U_t^k \geq \mu^*, A_t = k) \leq S(x) + \frac{2}{1 \wedge d(x \| \mu^k)},$$

where $S(x) \leq \frac{\gamma_T}{d(x \| \mu^*)}$.

It remains to control $\sum \mathbb{P}(L_t^* > \alpha) + \mathbb{P}(U_t^* < \mu^*)$. To control the second term, we first exploit the monotonicity of $d(\hat{\mu}_t^* \| q)$ on $[\hat{\mu}_t^*, 1]$ to note that

$$\{U_t^* > \mu^*\} = \{\max\{q > \hat{\mu}_t^* : d(\hat{\mu}_t^* \| q) \leq \gamma_t / N_t^*\} < \mu^*\} = \{\hat{\mu}_t^* < \mu^*, d(\hat{\mu}_t^* \| \mu^*) > \gamma_t / N_t^*\}.$$

The final event is the subject of (Theorem 10, Garivier & Cappé, 2011), who show that for any $z > 0$, and any k

$$\mathbb{P}(N_t^k d(\hat{\mu}_t^k \| \mu^k) > z) \leq e(z \log(t) + 1) e^{-z} \tag{7}$$

The statement extends, of course, to the empirical mean of any subsampling of any i.i.d. process in $[0, 1]$. The gist of the argument is to partition the space according to the size of N_t^k. If N_t^k is non-trivially large at some fixed time t, then it is exponentially unlikely for $N d(\hat{\mu}_t^k \| \mu^k)$ to exceed z, essentially because the cumulant generating function is bounded by that of a Bernoulli, and d is the Fenchel dual of this function for the Bernoulli. It is then just a question of stitching together these bounds over a well-chosen grid of values that N_t^k may take (concretely, a geometrically increasing grid is used, and we end up with a $\log t$ due to this grid), and accounting for the poor behaviour for small N_t^k (whence the premultiplying e). The argument presented in the supplement to the follow up work by Cappé et al. (2013) is somewhat cleaner than the original, and might be preferred.

Applying (7) to $\hat{\mu}_t^*$ and $z = \gamma_t$, we find that

$$\mathbb{P}(U_t^* < \mu^*) \leq e(\gamma_t \log(t) + 1) e^{-\gamma_t},$$

and so

$$\begin{aligned} \sum_{t=3}^{T} \mathbb{P}(U_t^* < \mu^*) &\leq \sum_{t=3}^{T} \frac{e(\log^2 t + 3 \log t \cdot \log\log t + 1}{t \log^3(t)} \\ &\leq e(\log\log T + 4). \end{aligned} \tag{8}$$

Control on $\sum \mathbb{P}(L_t^* > \alpha)$ follows identically. Exploiting monotonicity twice,

$$\{L_t^* > \alpha\} = \{\hat{\nu}_t^* > \alpha, d(\hat{\nu}_t^* \| \alpha) > \gamma_t / N_t^*\} \subset \{\hat{\nu}_t^* > \nu^*, d(\hat{\nu}_t^* \| \nu^*) > \gamma_t / N_t^*\},$$

and thus, applying (7) to $\hat{\nu}_t^*$ with $z = \gamma_t$,

$$\sum_{t=3}^{T} \mathbb{P}(k^* \notin \Pi_t) = \sum_{t=3}^{T} \mathbb{P}(L_t^* > \alpha) \leq e \log\log T + 4e. \tag{9}$$

Putting these together, we have

$$\mathbb{E}[N_t^k] \leq \frac{\log T + 3 \log\log T}{d(x \| \mu^*)} + 6 \log\log T + 24 + \frac{2}{1 \wedge d(x \| \mu^k)},$$

where we have used $2e < 6, 8e + 2 < 24$. □

C Proofs for Thompson Sampling with Optimistic Safety Indices

The first observation is that since the safety index L_t^k remains unchanged, we may directly use the proofs of Lemma 7 to observe that the bounds (4) and (9) continue to hold, that is,

$$\mathbb{E}[N_T^k] \leq \inf_y \frac{1}{\mathbb{1}\{\alpha < y < \nu^k\}} \left(\frac{\log T + 3\log\log T}{d(y\|\alpha)} + \frac{2}{1 \wedge d(y\|\nu^k)} \right),$$

$$\sum_{t=3}^{T} \mathbb{P}(k^* \notin \Pi_t) \leq e \log\log T + 4e.$$

The focus of the study then is to ensure that the TS analysis extends to control the play of inefficient arms. This pretty much exploits the analysis of TS due to Agrawal & Goyal (2013), although alternate analyses such as that of Kaufmann et al. (2012b) can equivalently be used.

The main bound is summarised in the following

Lemma 9 (Adaptation of Agrawal & Goyal (2013))**.** *There exists a universal constant C such that if $\Delta^k > 0$, then for any u, v such that $\mu^k < u < v < \mu^*$,*

$$\sum_{t=1}^{T} \mathbb{P}(A_t = k, k^* \in \Pi_t) \leq \frac{\log T}{d(u\|v)} + \frac{3}{1 \wedge d(u\|\mu^k)} + \frac{C}{(\mu^* - v)^2} \left(1 + \log \frac{1}{\mu^* - v} + \log \left(\frac{1}{1 - e^{-d(v\|\mu^*)}} \wedge T(\mu^* - v) \right) \right) \tag{10}$$

Let us first demonstrate the result from the main text using the above Lemma.

Proof of Theorem 3. We first argue the theorem.

For infeasible arms, instantiate (4) with a y such that $d(y\|\alpha) = d(\nu^k\|\alpha)/(1+\varepsilon)$. Since as previously argued, the resulting $d(y\|\nu^k)$ is $\Theta(\varepsilon^2)$.

For inefficient arms, consider the decomposition

$$\mathbb{E}[N_T^k] = \sum_{t=1}^{T} \mathbb{P}(A_t = k) \leq \sum_{t=1}^{T} \mathbb{P}(k^* \notin \Pi_t) + \sum_{t=1}^{T} \mathbb{P}(k^* \in \Pi_t, A_t = k).$$

The first term is bounded as $3\log\log T$. For the second term, we instantiate the bound (10) with a u and a v chosen so that

1. $d(u\|\mu^*) = d(\mu^k\|\mu^*)/\sqrt{1+\varepsilon}$
2. $d(u\|v) = d(u\|\mu^*)/\sqrt{1+\varepsilon} = d(\mu^k\|\mu^*)/(1+\varepsilon)$,

both of which exist by continuity.

Showing the bound then requires control on $u - \mu^k$ and $\mu^* - v$ (using the upper bound $d(a\|b) \geq 2(a-b)^2$). To this end, as in the proof of Theorem 1, observe that $u = \mu^k + \Theta(\sqrt{1+\varepsilon} - 1) = \mu^k + \Theta(\varepsilon)$. Similarly, $v = \mu^* - \Theta(\varepsilon)$. Therefore, $d(u\|\mu^k), d(v\|\mu^*) = \Theta(\varepsilon^{-2})$. Finally, since this ε^{-2} term does not grow with T, $(d(v\|\mu^*))^{-1} \wedge T = O(\varepsilon^{-2})$.

We may now conclude the argument exactly as in the proof of Theorem 1 ∎

Similarly to the case for Algorithm 1, this scheme also admits a gap-independent bound.

Proposition 10. *Algorithm 2, instantiated with* KL-UCB *type lower confidence bounds, attains the gap independent regret bound*

$$\mathbb{E}[\mathcal{R}_T] \leq O(\sqrt{KT\log T} + K\log\log T).$$

Proof. For infeasible arms, instantiate (4) with $y = (\alpha + \nu^k)/2$ to conclude that

$$\mathbb{E}[N_T^k] \leq O\left(\frac{\log T}{(\Gamma^k)^2}\right)$$

For inefficient arms, instantiate (10) with $u = \mu^k + \Delta^k/3$, and $v = \mu^k + 2\Delta^k/3$. Then $\mu^* - v = v - u = u - \mu^k = \Delta^k/3$, and by observing that $d(v\|\mu^*)^{-1} \wedge T\Delta^k/3 \leq T\Delta^k/3$, we have the upper bound

$$\mathbb{E}[N_T^k] \leq O(\log\log T) + O\left(\frac{\log T}{(\Delta^k)^2} + \frac{1 + \log(1/\Delta^k) + \log(T\Delta^k)}{(\Delta^k)^2}\right) = O\left(\log\log T + \frac{\log T}{(\Delta^k)^2}\right).$$

Taking the tighter of these bounds, and partitioning according to the size of $\Delta^k \vee \Gamma^k$, we have the bound

$$\mathbb{E}[\mathcal{R}_T] \leq \inf_{\mathbf{M}>0} T\mathbf{M} + O\left(\frac{K\log T}{\mathbf{M}}\right) + O(K\log\log T),$$

giving the claim upon optimisation. ∎

It remains to show the key Lemma. Again, we note that the key ideas are due to Agrawal & Goyal (2013).

Proof of Lemma 9. Fix a k. The values u and v essentially represent indices that we can compare the random scores ρ_t^* and ρ_t^k to. To this end, we define the 'good' events

$$\mathcal{G}_t^{\mu,k} := \{\widehat{\mu}_t^k \leq u\},$$
$$\mathcal{G}_t^{\rho,k} := \{\rho_t^k \leq v\}.$$

Notice that $\mathcal{G}_t^{\mu,k}$ lies in $\mathscr{H}_{t-1}$.

Now, we start with the decomposition

$$\begin{aligned}\mathbb{P}(A_t = k, k^* \in \Pi_t) &= \mathbb{P}(A_t = k, k^* \in \Pi_t, \mathcal{G}_t^{\mu,k}, \mathcal{G}_t^{\rho,k}) + \mathbb{P}(A_t = k, k^* \in \Pi_t, \mathcal{G}_t^{\mu,k}, (\mathcal{G}_t^{\rho,k})^c) + \mathbb{P}(A_t = k, k^* \in \Pi_t, (\mathcal{G}_t^{\mu,k})^c) \\ &\leq \mathbb{P}(A_t = k, k^* \in \Pi_t, \mathcal{G}_t^{\mu,k}, \mathcal{G}_t^{\rho,k}) + \mathbb{P}(A_t = k, \mathcal{G}_t^{\mu,k}, (\mathcal{G}_t^{\rho,k})^c) + \mathbb{P}(A_t = k, (\mathcal{G}_t^{\mu,k})^c). \qquad (11)\end{aligned}$$

Now, the last of these terms in (11) is easily controlled - indeed, $\mathbb{P}(A_t = k, (\mathcal{G}_t^{\mu,k})^c) = \mathbb{P}(A_t = k, \widehat{\mu}_t^k > u)$ is exponentially small if N_t^k is large. In fact, mirroring the approach of the proof of Lemma 7, we find that

$$\begin{aligned}\sum_{t\leq T} \mathbb{1}\{A_t = k, \widehat{\mu}_t^k > u\} &= \sum_{t\leq T}\sum_{s\leq t} \mathbb{1}\{A_t = k, N_t^k = s, \widehat{\mu}_t^k > u\} \\ &= \sum_s \mathbb{1}\{\hat{\mu}^k(s) > u\} \sum_{t\geq s} \mathbb{1}\{A_t = k, N_t^k = s\} \\ &\leq \sum_{s\leq T} \mathbb{1}\{\hat{\mu}^k(s) > u\},\end{aligned}$$

where we set $\hat{\mu}^k(s)$ to be the value of $\widehat{\mu}_t^k$ at the first t such that $N_t^k = s$. But then, by Chernoff's bound, $P(\hat{\mu}^k(s) > u) \leq \exp(-sd(u\|\mu^k)$, giving the bound

$$\sum \mathbb{P}(A_t = k, (\mathcal{G}_t^{\mu,k})^c) \leq \frac{2}{1 \wedge d(u\|\mu^k)}. \qquad (12)$$

The second term of (11) too is similar to control, upon observing that the posterior Beta law is very well concentrated around $\widehat{\mu}_t^k$ with variance scale $1/N_t^k$. More concretely, Agrawal & Goyal (2013) exploit the

following observation: if $F(x;\mathrm{Beta}(a,b))$ is the CDF of a $\mathrm{Beta}(a,b)$ random variable, and $G(k;\mathrm{Bin}(n,p))$ is the CDF of a Binomial random variable, then for natural $n \geq k$,

$$1 - F(x;\mathrm{Beta}(k+1, n-k+1)) = G(k;\mathrm{Bin}(n+1,x)).$$

This relation most easily follows from the fact that the $\mathrm{Beta}(k+1,n-k+1)$ is the law of the $k+1$th order statistic of $n+1$ samples from the uniform distribution, and the chance of this exceeding x is simply the chance that the k smaller ones are at most x, and the rest are at least x, which of course is expressed by the Binomial distribution. But then we conclude that for any N_0

$$\mathbb{P}(\rho_t^k > v | N_t^k > N_0, \hat{\mu}_t^k \leq u) \leq e^{-N_0 d(v\|u)}.$$

Choosing $N_0 = \log(T)/d(v\|u)$, we then get the bound

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}(A_t = k, \rho_t^k > v, \widehat{\mu}_t^k \leq u) &\leq \sum_{t=1}^{T} \mathbb{P}(A_t = k, N_T^k \leq N_0) + \sum_{t=1}^{T} \mathbb{P}(N_T^k > N_0, \rho_t^k > v, \widehat{\mu}_t^k \leq u) \\
&\leq N_0 + T e^{-N_0 d(v\|u)} \\
&\leq \frac{\log T}{d(v\|u)} + 1 \leq \frac{\log T}{d(v\|u)} + \frac{1}{1 \wedge d(u\|\mu^k)}. \qquad (13)
\end{aligned}$$

This leaves the first term of (11), which is the hardest to control, and ultimately relies upon hard analysis of Binomial tails. The idea is roughly to use v as a lower index for ρ_t^*. Indeed, let

$$\mathbf{P}_t := \mathbb{P}(\rho_t^* > v | \mathscr{H}_{t-1}) = \mathbb{P}_{t-1}(\rho_t^* > v).$$

Then observe that

$$\begin{aligned}
\mathbb{P}_{t-1}(A_t = k, \mathcal{G}_t^{\mu,k}, \mathcal{G}_t^{\rho,k}, k^* \in \Pi_t) &= \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} \mathbb{P}_{t-1}(A_t = k, \rho_t^k < v) \\
&\leq \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} \mathbb{P}_{t-1}(\forall k \in \Pi_t, \rho_t^k < v) \\
&= \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} (1 - \mathbf{P}_t) \mathbb{P}_{t-1}(\forall k \neq k^* \in \Pi_t, \rho_t^k < v) \\
&= \frac{1 - \mathbf{P}_t}{\mathbf{P}_t} \mathbb{1}\{\mathcal{G}_t^{\mu,k}, k^* \in \Pi_t\} \mathbb{P}_{t-1}(\rho_t^* > v, \forall k \neq k^* \in \Pi_t, \rho_t^k < v) \\
&\leq \frac{1 - \mathbf{P}_t}{\mathbf{P}_t} \mathbb{P}_{t-1}(A_t = k^*),
\end{aligned}$$

where we have used the fact that $\mathcal{G}_t^{\mu,k} \in \mathscr{H}_{t-1}$ and Π_t is predictable. The idea is to now exploit the fact that $\mathbf{P}_t$ is exponentially close to 1 as N_t^* increases, and by expressing this chance in terms of the size of N_t^* and analysing the same, Agrawal & Goyal (2013) show in their Lemma 2 that

$$\sum_{t=1}^{T} \mathbb{E}[(1 - \mathbf{P}_t)\mathbb{P}_{t-1}(A_t = k^*)/\mathbf{P}_t] \leq \frac{24}{\Delta_v^2} + C' \sum_{s \geq 8/\Delta_v}^{T-1} e^{-\Delta_v^2 s/2} + \frac{1}{e^{\Delta_v^2 s/4} - 1} + \frac{e^{-s d(v\|\mu^*)}}{(s+1)\Delta_v^2},$$

where $\Delta_v := (\mu^* - v)$ and C' is a constant. Notice that each of the terms in the sum are monotonically decreasing. Therefore, we may derive upper bounds by comparison to an integral, which yields for the first and second terms that

$$\sum_{s=\lceil 8/\Delta_v \rceil}^{T-1} e^{-\Delta_v^2 s/2} \leq \int_0^{\infty} e^{-\Delta_v^2 s/2} \mathrm{d}s = \frac{2}{\Delta_v^2},$$

and

$$\begin{aligned}\sum_{s=\lceil 8/\Delta_v\rceil}^{T-1}\frac{1}{e^{\Delta_v^2 s/4}-1} &\le \int_{7/\Delta_v}^{T}\frac{1}{e^{\Delta_v^2 s/4}-1}\mathrm{d}s\\ &= \frac{4}{\Delta_v^2}\int_{(7/4\Delta_v}^{\Delta_v^2 T/4}\frac{1}{e^u-1}\mathrm{d}u\\ &\le \frac{4}{\Delta_v^2}\log\frac{1}{1-e^{-7/4\Delta_v}}\\ &\le \frac{4}{\Delta_v^2}\log\frac{2}{1\wedge 7/4\Delta_v} \le \frac{4}{\Delta_v^2}\left(\log\frac{1}{\Delta_v}+O(1)\right),\end{aligned}$$

where we have used the previously established fact that $\frac{1}{1-e^{-x}} \le \frac{2}{x\wedge 1}$.

For the final term, we may bound this in two ways - firstly simply observing that $e^{-sd} \le 1$, we get the bound $\sum_{8/\Delta_v}^{T-1}\frac{1}{s+1} \le \log(T\Delta/8)$. In addition, we derive a T-independent bound as follows, wherein we abbreviate $d_v = d(v\|\mu^*)$.

$$\begin{aligned}\sum_{s=\lceil 8/\Delta_v\rceil}^{T-1}\frac{e^{-sd_v}}{(s+1)\Delta_v^2} &= e^{d_v}\sum_{s=\lceil 8/\Delta_v}^{T-1}\frac{e^{-(s+1)d_v}}{s+1}\\ &= e^{d_v}\sum_{s=\lceil 8/\Delta_v\rceil}^{T-1}\int_{u=d_v}^{\infty}e^{-(s+1)u}\mathrm{d}u\\ &\le e^{d_v}\int_{u=d_v}^{\infty}\sum_{s=1}^{\infty}e^{-(s+1)u}\mathrm{d}u\\ &= e^{d_v}\int_{u=d_v}^{\infty}\frac{e^{-u}}{e^u-1}\mathrm{d}u\\ &\le \log\frac{1}{1-e^{-d_v}} \le \log\frac{2}{d_v}+O(1).\end{aligned}$$

Taking the smaller of these two bounds, the final term is controlled by $4\Delta_v^{-2}[\log(T\Delta_v\wedge d_v^{-1})+O(1)]$, and we have

$$\sum_{t=1}^{T}\mathbb{P}(A_t=k,\mathcal{G}_t^{\mu,k},\mathcal{G}_t^{\rho,k},k^*\in\Pi_t) \le \frac{C}{\Delta_v^2}\left(1+\log\frac{1}{\Delta_v}+\log\left(\Delta_v T\wedge -d(v\|\mu^*)^{-1}\right)\right). \tag{14}$$

The claimed bound is then realised by adding up (12, 13, 14). □

D Proofs for Thompson Sampling with BayesUCB

Since the procedure for selecting arms given Π_t is left unchanged from the previous case, we only need to demonstrate that Π_t is good, that is, that the lower bound index L_t^k performs well. Indeed, this is essentially exploiting the fact that the argument of the previous section only uses the fact that Π_t is a predictable process, and then specifics of the Thompson scores ρ_t^ks, and so the second term of the decomposition

$$\sum_t\mathbb{P}(A_t=k) \le \sum_t\mathbb{P}(k^*\notin\Pi_t)+\sum_t\mathbb{P}(k^*\in\Pi_t,A_t=k)$$

can be pursued identically to control the play of inefficient arms on rounds such that $k^* \in \Pi_t$, again giving (10).

We show the following bound, following the methods of Kaufmann et al. (2012a) as described in §A.

Lemma 11. *In the setting of Theorem 4, the following hold.*

- *If* $\Gamma^k > 0$, *then for any* $x \in (\alpha, \nu^k)$,

$$\mathbb{E}[N_T^k] \leq \frac{3/2 \log T + 3 \log\log T + 3/2 \log 2}{d(x\|\alpha)} + \frac{2}{1 \wedge d(x\|\nu^k)} \tag{15}$$

- *The mean number of times the optimal arm is treated as impermissible is bounded as*

$$\sum_{t=3}^{T} \mathbb{P}(k^* \notin \Pi_t) \leq e \log\log T + 4e.$$

The claimed bound is quickly forthcoming upon combining the appropriate pieces of the proofs of Theorems 1 and 3.

Proof of Theorem 4. For inefficient arms, combining the second part of Lemma 11 and (10), we conclude that if $\Delta^k > 0$, then

$$\mathbb{E}[N_T^k] \leq \frac{\log T}{d(u\|v)} + \frac{3}{1 \wedge d(u\|\mu^k)} + \frac{C}{(\mu^* - v)^2}(1 + (d(v\|\mu^*)^{-1} \wedge \log T)) + e \log\log T + 4e.$$

Similarly, for infeasible arms, by using (15), we have the control

$$\mathbb{E}[N_t^k] \leq \frac{\log T + 3\log\log T + 2\log 2}{2/3 d(y\|\alpha)} + \frac{2}{1 \wedge d(y\|\nu^k)}.$$

Now choosing u, v, y as in the proof of Theorem 3 and proceeding along the same lines gives the claim. □

The same approach also shows the following gap-independent result. The proof is identical, and so omitted.

Proposition 12. *Algorithm 3 instantiated with* BayesUCB *with* $\delta_t^k = 1/\sqrt{8 N_t^k t \log^3 t}$ *also satisfies the bound*

$$\mathbb{E}[\mathcal{R}_T] = O(\sqrt{KT \log T} + K \log\log T).$$

We conclude by showing the main Lemma.

Proof of Lemma 11. The argument relies on the following estimate, which essentially serves as a reduction to the analysis of KL-UCB. This result is a variation of Lemma 1 of Kaufmann et al. (2012a).

Lemma 13. *Define the quantities*

$$\underline{\varphi}_t^k := \mathbb{1}\{S_t^k > 0\} \min\left\{ q \leq \frac{S_t^k}{N_t^k} : N_t^k d\left(\frac{S_t^k}{N_t^k} \middle\| q\right) \leq \log((2t \log^2 t)^{3/2}) \right\}$$
$$\overline{\varphi}_t^k := \mathbb{1}\{S_t^k > 0\} \min\left\{ q \leq \frac{S_t^k}{N_t^k} : N_t^k d\left(\frac{S_t^k}{N_t^k} \middle\| q\right) \leq \log(t \log^3(t)) \right\}.$$

Then for all t,

$$\underline{\varphi}_t^k \leq L_t^k \leq \overline{\varphi}_t^k.$$

Proof. Firstly, since $L_t^k = 0$ whenever $S_t^k = 0$, this case is trivial. So assume $S_t^k \geq 1$.

The idea behind the bounds is to exploit the relationship between the CDFs of Beta and Binomial random variables to reduce the quantile estimation to that of a Binomial, and then use Chernoff's bound for the Binomial to control where the quantile can be. Indeed, let $Z \sim \text{Beta}(S_t^k, N_t^k - S_t^k + 1)$. Then we know that

$$\mathbb{P}(Z \leq q) = \mathbb{P}(\text{Bin}(N_t^k, q) \geq S_t^k).$$

Further, by Chernoff's upper bound, and by estimating the sth term in the Binomial series using Stirling's approximation, we may show the following result (where the lower bound holds generally, and the upper bound holds for any $s \geq nq$).

$$\frac{1}{\sqrt{8n}} \exp(-nd((s/n)\|q)) \leq \mathbb{P}(\text{Bin}(n, q) \geq s) \leq \exp(-nd((s/n)\|q)).$$

Now, recall that L_t^k is the δ_t^kth quantile of the law of Z, so that $P(Z \leq L_t^k) = \delta_t^k$.

Lower bound Suppose $q \leq S_t^k/N_t^k$ is such that

$$\exp(-N_t^k d(S_t^k/N_t^k)\|q)) \leq \delta_t^k.$$

Then it follows that $q \leq L_t^k$. Therefore,

$$\begin{aligned} L_t^k &\geq \max\left\{q \leq \frac{S_t^k}{N_t^k} : N_t^k d\left(\frac{S_t^k}{N_t^k}\|q\right) \geq \log(1/\delta_t^k)\right\} \\ &= \min\left\{q \leq \frac{S_t^k}{N_t^k} : N_t^k d\left(\frac{S_t^k}{N_t^k}\|q\right) \leq \log(1/\delta_t^k)\right\}, \end{aligned}$$

where the final equality is due to the continuity of $d(a\|\cdot)$.

Now observe that

$$\log(1/\delta_t^k) \leq (2(t+1))^{3/2} \log^3 t.$$

Therefore, replacing $\log(1/\delta_t^k)$ by the larger $\log(2(t+1)^{3/2} \log^3 t$ in the lower bound can only decrease it.

Upper bound Suppose that $q \leq S_t^k/N_t^k$ is such that the lower bound on the Binomial tail exceeds δ_t^k. Then L_t^k must be smaller than this q, and so

$$L_t^k \leq \min\left\{q \leq \frac{S_t^k}{N_t^k} : N_t^k d\left(\frac{S_t^k}{N_t^k}\|q\right) \leq \log\left(\frac{1}{\sqrt{8N_t^k}\delta_t^k}\right)\right\}.$$

But, by definition,

$$\frac{1}{\sqrt{8N_t^k}\delta_t^k} = t \log^3 t.$$

□

Observe that the bounds $\overline{\varphi}$ and $\underline{\varphi}$ exactly take the form of the KL-UCB bounds, but with a different value for γ_T. Thus, the same proofs may be repeated.

Indeed, to show (15), we observe that for an arm with a safety gap, $\{L_t^k \leq \alpha\} \subset \{\underline{\varphi}_t^k \leq \alpha, \}$ and we may then follow the proof of Lemma 7 to control this identically to there - the only change is that $\log(\gamma_T)$ in $S(y)$ is replaced by $\log((2t \log^2 t)^{3/2})$.

Further, the upper bound is exactly the bound of KL-UCB, and therefore without alteration we may immediately conclude that

$$\sum_{t\geq 3}\mathbb{P}(L_t^* > \alpha) \leq \sum_{t\geq 3}\mathbb{P}(\overline{\varphi}_t^* > \alpha) \leq e\log\log t + 4e. \qquad \square$$

We note that the last property in the proof of Lemma 13 is exactly the reason for selecting δ_t of the form that we did, which is essentially the $1/\gamma_t$ from KL-UCB, but scaled down to ensure that the BayesUCB bound is at least as optimistic as that of KL-UCB. In principle, then, this gives an avenue for a tighter analysis by choosing a more refined notion of δ_t by exploiting stronger bounds for the Binomial tails.

For instance, it is known (Prop A.4, A.2 Jeřábek, 2004) that there exists a constant C such that for $s \geq nq + \sqrt{nq(1-q)}$,

$$\frac{1}{C}\frac{qn-qs}{s-qn}\sqrt{\frac{n}{s(n-s)}}e^{-nd(s/n\|q)} \leq \mathbb{P}(\mathrm{Bin}(n,q)\geq s) \leq C\frac{qn-qs}{s-qn}\sqrt{\frac{n}{s(n-s)}}e^{-nd(s/n\|q)},$$

while for $s \leq nq + \sqrt{nq(1-q)}$, it is bounded below by another constant C'. This suggests using $\delta_t \sim \min\left(C', \frac{1}{t\log^3 t}\cdot\sqrt{\frac{N_t^k}{S_t^k(N_t^k-S_t^k)}}\right)$, although it is unclear how to handle the $(qn-qs)/(s-qn)$ term properly. Assuming this is indeed handled, though, this should result in an improvement to $\overline{\varphi}$ of replacing the $t^{3/2}$ by something $O(t)$, while the lower bound should remain unchanged. Of course, this does not quite explain the success of $\delta_t = 1/t$ in the experiments, and it is possible that this approach simply serves to make BayesUCB look more like KL-UCB, which defeats the purpose somewhat.

E Lower Bound

We begin by showing the key Lemma.

Proof of Lemma 5. Fix a (possibly randomised) algorithm. Let $\{\mathbb{P}^k\}$ and $\{\widetilde{\mathbb{P}}^k\}$ be two safe bandit instances, and recall that $\mathscr{H}_t := \{(A_s, R_s, S_s) : s \leq t\}$ denotes the history of play. We will use $\mathbb{P}$ to represent laws in the first instance and $\widetilde{\mathbb{P}}$ for laws in the second. Similary, $\mathbb{E}$ and $\widetilde{\mathbb{E}}$ denote expectations under the two laws.

Let Z be any function of measurable with respect to $\sigma(\mathscr{H}_T)$ that is bounded in $[0,1]$. Then observe that from $\mathscr{H}_T$, we can generate a random bit by first computing $Z(\mathscr{H}_{T+1})$, and then sampling $B \sim \mathrm{Bern}(Z)$. Clearly, the mean of B is the same as that of Z. But then, by the data processing inequality,

$$D(\mathbb{P}_{\mathscr{H}_T}\|\widetilde{\mathbb{P}}_{\mathscr{H}_T}) \geq D(\mathbb{P}_B\|\widetilde{\mathbb{P}}_B) = d(\mathbb{E}[Z]\|\widetilde{\mathbb{E}}[Z]).$$

Next, due to the chain rule of KL divergence, for any $t \geq 1$,

$$\begin{aligned} D(\mathbb{P}_{\mathscr{H}_t}\|\widetilde{\mathbb{P}}_{\mathscr{H}_t}) &= D(\mathbb{P}_{\mathscr{H}_{t-1}}\|\widetilde{\mathbb{P}}_{\mathscr{H}_{t-1}}) \\ &\quad + \mathbb{E}[D(\mathbb{P}_{A_t|\mathscr{H}_{t-1}}\|\widetilde{\mathbb{P}}_{A_t|\mathscr{H}_{t-1}}|\mathscr{H}_{t-1})] \\ &\quad + \mathbb{E}[D(\mathbb{P}_{(R_t,S_t)|A_t,\mathscr{H}_{t-1}}\|\widetilde{\mathbb{P}}_{(R_t,S_t)|A_t,\mathscr{H}_{t-1}}|A_t,\mathscr{H}_{t-1})]. \end{aligned}$$

Now, the second term in the RHS is 0 since the learner must be causal, and thus the law of A_t is determined by $\mathscr{H}_{t-1}$. Further, the feedback (R_t, S_t) is independent of the history given A_t, and is distributed according to $\mathbb{P}^{A_t}$ and $\widetilde{\mathbb{P}}^{A_t}$ under the two instances. We thus have the recurrence

$$D(\mathbb{P}_{\mathscr{H}_t}\|\widetilde{\mathbb{P}}_{\mathscr{H}_t}) - D(\mathbb{P}_{\mathscr{H}_{t-1}}\|\widetilde{\mathbb{P}}_{\mathscr{H}_{t-1}}) = \sum_k \mathbb{P}(A_t = k)D(\mathbb{P}^k\|\widetilde{\mathbb{P}}^k).$$

Summing this up, and observing that $\mathscr{H}_0$ is trivial, and then recalling $\sum_t \mathbb{P}(A_t = k) = \mathbb{E}[N_T^k]$, it follows that

$$D(\mathbb{P}_{\mathscr{H}_T} \| \widetilde{\mathbb{P}}_{\mathscr{H}_T}) = \sum_k \mathbb{E}[N_T^k] D(\mathbb{P}^k \| \widetilde{\mathbb{P}}^k).$$

The conclusion now follows on taking $Z = N_T^k/T$, which trivially lies in $[0,1]$. □

Proof of Proposition 6. As mentioned in the main text, choose $\widetilde{\mathbb{P}}^j = \mathbb{P}^j$ for $j \neq k$, and instead let $\widetilde{\mathbb{P}}^k$ be any law on $\{0,1\}^2$ of means $(\mu^k \vee \mu^* + \varepsilon, \nu^k \wedge \alpha)$. Notice that in the $\widetilde{\mathbb{P}}$-instance, arm k is optimal.

Since the algorithm ensures that suboptimal arms are not played more than $C_x T^x$ times, $\mathbb{E}[N_T^k/T] \leq C_x T^{-(1-x)}$, and $\widetilde{\mathbb{E}}[N_T^k/T] \geq 1 - C_x T^{-(1-x)}$ for any $x \in (0,1)$. Therefore,

$$\begin{aligned} d(\mathbb{E}[N_T^k/T] \| \widetilde{\mathbb{E}}[N_T^k/T]) &\geq \left(1 - \frac{\mathbb{E}[N_T^k]}{T}\right) \log \frac{1}{1 - \widetilde{\mathbb{E}}[N_T^k/T]} - \log 2 \\ &\geq (1 - o(1))(1-x) \log \frac{T}{C_x} - \log 2 = (1 - o(1))(1-x) \log T. \end{aligned}$$

Next, since we are working with independent means and safety rewards, taking $\widetilde{\mathbb{P}}^k$ to also have the independent rewards, we get $D(\mathbb{P}^k \| \widetilde{\mathbb{P}}^k) = d_<(\mu^k \| \mu^* + \varepsilon) + d_>(\nu^k \| \alpha)$.

We conclude that for any $x, \varepsilon \in (0,1)$,

$$\frac{\mathbb{E}[N_T^k]}{\log T} \geq \frac{(1-x)(1-o(1))}{d_<(\mu^k \| \mu^* + \varepsilon) + d_>(\nu^k \| \alpha)},$$

whence the claim follows on taking $\underline{\lim}_{T \nearrow \infty}$, and then taking limits as $x \to 0, \varepsilon \to 0$, and exploiting the continuity of $d_<(a\|b)$. □

F Simulation Details and Supplementary Plots

Implementation Details All methods are implemented on MATLAB. Throughout we use independent Bernoulli bits for both R and S. The particular details of the methods used are described below.

Policy approaches It is a straightforward observation that for a single constraint and objective, the solution to linear program $\max_{\pi \in \Delta} \langle \pi, a \rangle$ s.t. $\langle \pi, b \rangle \leq c$ is supported on at most two coordinates. Further, the optimal policy on two given coordinates itself is simple to compute - clearly at least one needs to be safe according to the relevant safety index at the particular time, else this is not a permitted policy. If both are safe as per the index, then the policy can concentrate on the one with larger reward index. If one is safe and the other not, then the policy concentrates on the safe one if it has a larger reward index. Otherwise, we assign the slack between the safety level and the safety index of the safe coordinate as the mass of the policy on the coordinate with the unsafe index. This enables a simple - and fast - method to select the round-wise policies for both BwCR and Pess - we simply evaluate the value of the optimal policy on each pair of arms, and choose the one with the largest reward.

Details of Confidence Bound Computation

In effect we use two types of confidence bounds - KL-UCB-based, and BayesUCB-based.

- KL-UCB-based bounds are all evaluated with $\gamma_t = 1/t$ (i.e., without the extra $1/\log^3 t$ factor in the main text). This is aligned with the practical recommendations of Garivier & Cappé (2011).

  The upper indices U_t^k on μ^k are computed simply by computing a lower bound for $1 - \mu^k$, and then subtracting this from one. The soundness of this procedure is a trivial exercise.

Finally, the KL inversion is performed via a binary search. Specifically, we carry this out for $\max(4, \log_2(t))$ rounds, thus ensuring that any error in the estimate is of order $1/t$, which ensures that extra regret due to numerical precision is at most $\log T$.

- BayesUCB -based bounds are all evaluated with $\delta_t^k = 1/(t+1)$. Again, this is in line with the recommendations of Kaufmann et al. (2012a). We note that this is a larger quantile than studied in the main text, and a regret bound with this δ_t^k is currently unavailable. Nevertheless, the empirical performance is sound, as seen in §6.

  The quantile estimation is performed by using the library `betainv` function provided by the Statistics Toolbox of MATLAB. This uses Newton's method to solve the equation defining a quantile of a Beta distribution.

- For TS, we sample from the appropriate Beta posteriors by using the library `betarnd` function provided by MATLAB.

F.1 Supplement to §6.1

We provide plots that detail the regrets achieved by each algorithm in the two cases studied. The main observations remain unchanged - the regrets of policy based methods grow linearly in the first case, and while they appear sublinear inthe second, they are at a much larger scale than our implementation. We note that in both cases the more unsafe BwCR performs better on the regret criterion. This should be evident on the data of case two, for which playing the unsafe arm only contributes $0.6 - 0.5 = 0.1$ to the regret, while the suboptimal arm has a gap of $0.6 - 0.4 = 0.2$. However, the data of case 1 suggests that this is also true more broadly, and may be an effect of the optimism principle. That said, this is a moot point in this case since hte growth rate is very much linear.

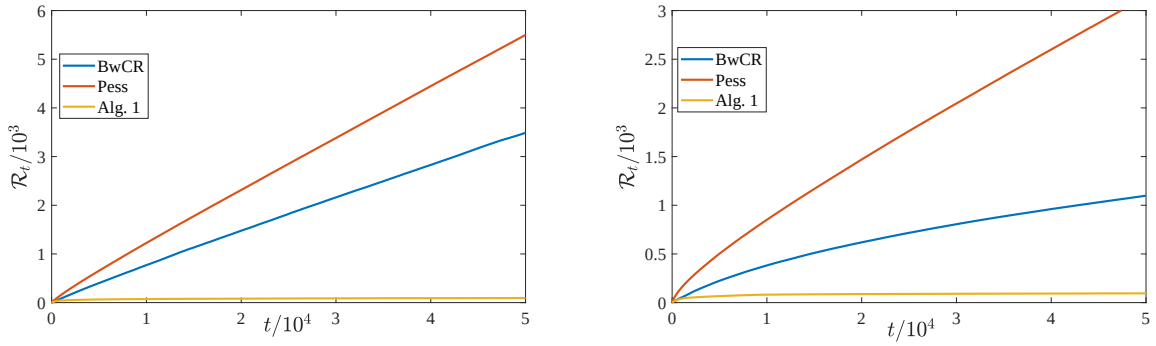

Figure 4: Regrets for the situations of §6.1 - left is the first case with two optimal policies, right is the second case with a single optimal policy supported on a single arm

F.2 Supplement to §6.2

We first provide Box plots in Figure 5 of the spread of regret and safety for the situation studied in the main text with $\alpha = 0.21$. Note that the Regret of the TS based methods shows somewhat larger fluctuations, although the maximum of the data is similar. For the net safety violation, the fluctuations are similarly sized, and the Bayesian methods retain an advantage.

Next, we provide plots for the same scenario, but with $\alpha = 0.19$. Note that this induces the difference that the arm 4 is now unsafe by a significant amount, which increases its gap $\Delta^4 \vee \Gamma^4$ to about 0.02 from 0.004. However, since 0.004 is about the same size as $\sqrt{K/T} = 0.01$, this arm was not contributing much to the regret in the previous case. Further, the safety violation of the least unsafe arm is not only about 0.02

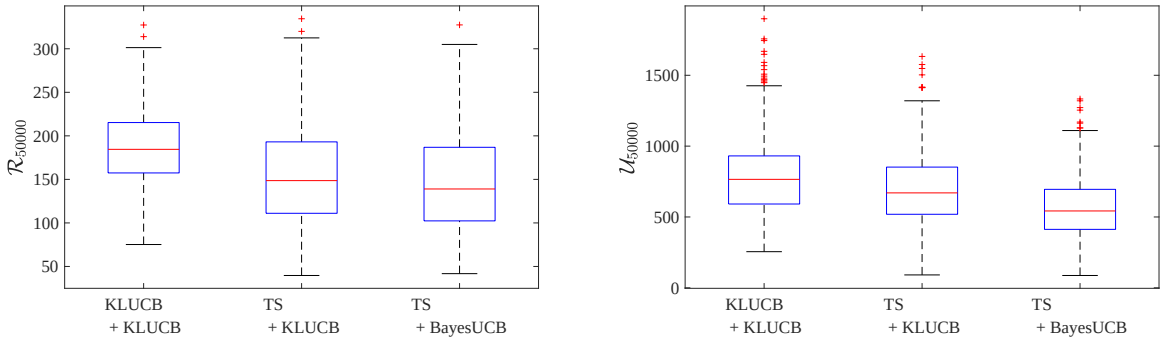

Figure 5: Box plots of Regret and Total Safety Violation at time $T = 50K$ over 500 runs for the Trial Drugs data with $\alpha = 0.21$

instead of the previous 0.05. Correspondingly, we expect to see an increase in the play of unsafe arms, as well as a slight increase in regret due to the scale up from 0.04 to 0.019 in the play of this arm. Both of these observations are clearly borne out in Figure 6, which presents data over 100 trials.

We note that these observations are again consistent with the theoretical bounds. The main term of the regret bound is roughly $40 \log t$, while that of the safety violation bound is roughly $1500 \log t$, and $\log(10^4) \approx 4 \cdot 2.3 \approx 10$.

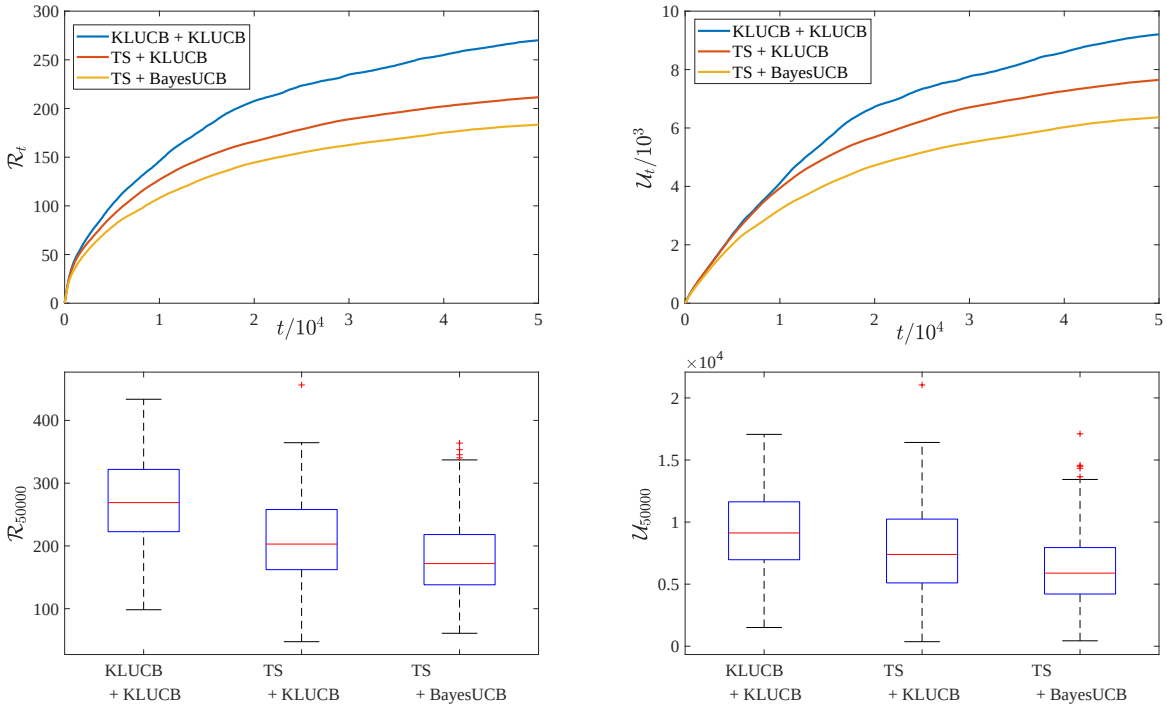

Figure 6: Top: Mean regret (left) and safety (right) violation as a function of t, averaged over 100 trials, for the Trial Drug data with $\alpha = 0.19$ Bottom: Box plot of the same at $T = 50K$.

F.3 Comparing theoretically analysed BayesUCB quantiles with the practically implemented ones

As observed in the main text, the simulation of §6 all present Algorithm 3 run with the quantile schedule $\delta_t^k = 1/t$ - in actuality, we use the slightly more reasonable schedule of $\min(\alpha/2, t^{-1})$, simply to ensure that for small t, all arms are declared as feasible. While this choice is consistent with the recommendation of Kaufmann et al. (2012a), it differs from the schedule analysed theoretically in §4.2, which instead suggsets $\delta_t^k = (\sqrt{8N_t^k t \log^3 t})^{-1})$. We present the behaviour of such a schedule below, although we modify it slightly to $\min(\alpha/2, (\sqrt{8N_t^k t})^{-1})$ - here we drop the log term as recommended by Garivier & Cappé (2011), and introduce the minimum to again ensure that for small t all arms are declared to be feasible. The resulting behaviour is compared with the previously studied $1/t$ schedule in Figure 7 on the simple data $\mu = \nu = (0.4, 0.5, 0.6), \alpha = 0.5$.

Observe that the theoretical schedule displays the favourable logarithmic growth, and so is consistent with Theorem 4. Further, while it certainly suffers degradation relative to the $1/t$ schedule, this is limited. The reason for this degradation is largely because the theoretical lower indices L_t^k are more optimistic, and allow the unsafe arm to be played for a larger number of times, as borne out in the plot of total safety violations.

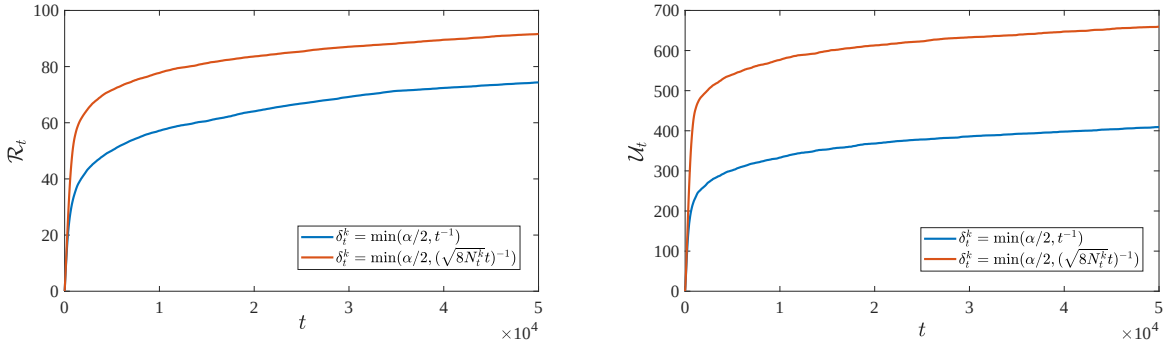

Figure 7: Regret (left) and total safety violations (right) of the theoretical and the $1/t$ schedule for Algorithm 3. Averages over 500 trials are presented.

Additionally, Figure 8 presents boxplots of the regret and net safety at $T = 50000$ for the two schedules. An interesting observation is that the schedule $1/t$ exhibits greater variability, with some (rare) but massive outliers that are not present for the theoretical schedule. Investigating this more closely requires determining high-probability bounds on these methods, which is a subject for future work.

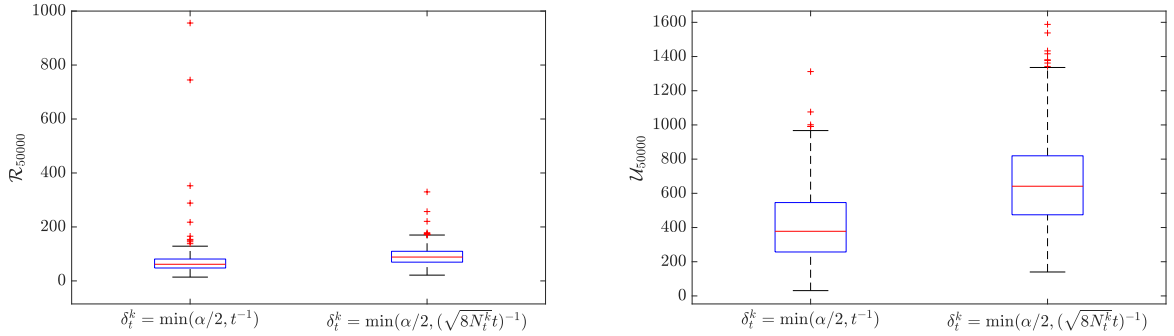

Figure 8: Boxplots across 500 trials of the regret (left) and safety violations (right) for the two schedules at $T = 50000$. One outlier for regret for the $1/t$ schedule at ≈ 2100 has been omitted for the sake of clarity.

F.4 The behaviour of a Naïve Thompson Sampling Based Safety Index

As discussed in §4.2, a naïve way of constructing a safety index by just sampling $\theta_t^k \sim \text{Beta}(S_t^k+1, N_t^k-S_t^k+1)$ should be ineffective when the safety score ν^* is close to α. We first investigate this effect.

Concretely, the scheme is the same as Alg.3, except that instead of the BayesUCB index, we construct a safety index by sampling as above, and then populate $\Pi_t = \{k : \theta_t^k \leq \alpha\}$. We run this scheme with the data

$$\mu = (0.3, 0.5, 0.7),$$
$$\nu = (0.3, 0.5, 0.7),$$

and *vary* α as $0.5 + i/50$ for $i \in [0:9]$. This corresponds to an increasing safety slack, while for $i \in [0:5]$, the safety gap of the unsafe arm 3 remains large, but decaying. Note that this ostensibly should increase the large t regret for a scheme with optimal dependence.

Figure 9 plots the resulting mean regrets over a horizon of length 10K for four of the 10 cases (chosen evenly to not clutter the plot too much). The data is averaged over 200 trials. Observe that for $i = 0$, wherein the gap $(\alpha - \nu^*)$ is 0, the regret grows linearly, while the dependence becomes sublinear as i increases, and further improves, even though it should grow like $1/i$.

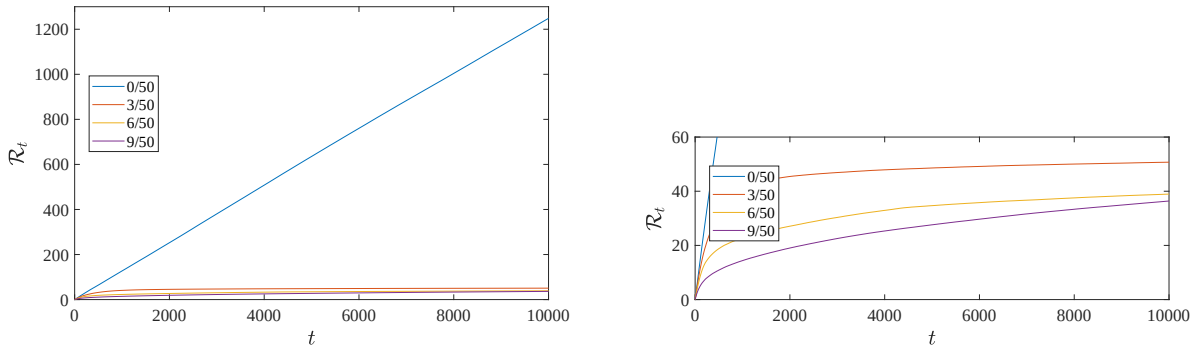

Figure 9: Regret of the scheme with a Naïve TS based safety index for various choices of α. The legend marks $\alpha - \nu^*$. The right figure zooms in to the bottom of the left figure.

Further, we observe that the dependence on $\alpha - \nu^*$ scales roughly as inverse-quadratic. This is illustrated in Figure 10, which plots both the mean regret against $\alpha - \nu^*$, as well as the mean of $1/\sqrt{(\mathcal{R}_t)}$ against $\alpha - \nu^*$. The key observation is the nearly linear dependence in the second plot for small $\alpha - \nu^*$. This observation makes sense - the variance scale of a $\text{Beta}(S+1, N-S+1)$ distribution is as $1/N$, and so if the means $\hat{\nu}^*$ is close to the truth, then the chance of θ_t^k falling above α at time t is roughly $1/t(\alpha - \nu^*)^2$, and so $k^* \notin \Pi_t$ for about $\log(T)/(\alpha - \nu^*)^2$ rounds. Of course, for large enough $\alpha - \nu^*$, this term is dominated by the regret terms due to suboptimal arms, and the dependence is masked. This effect is further confounded in our simulation with the fact that the safety gap Γ^3 reduces as α is increased, which raises the regret. Nevertheless, the trend is evident, at least in the low $\alpha - \nu^*$ regime where the gap Γ^3 does not change as much, and remains much larger than $\alpha - \nu^*$.

Despite the ineffectiveness when $\alpha - \nu^*$ is small, a TS based safety index is an attractive proposition, primarily due to wider concerns - the advantage of TS for standard bandits is obtaining strong regret performance at a low computational cost, and this is specially important in cases such as combinatorial or continuously armed bandits. An alternative sampling based strategy would enable such an approach for safe bandits in such rich scenarios, and is of both practical and theoretical interest. Promisingly, when the gap is large, the effect on regret is indeed mild, showing that this is the only obstacle in the path of such a strategy.

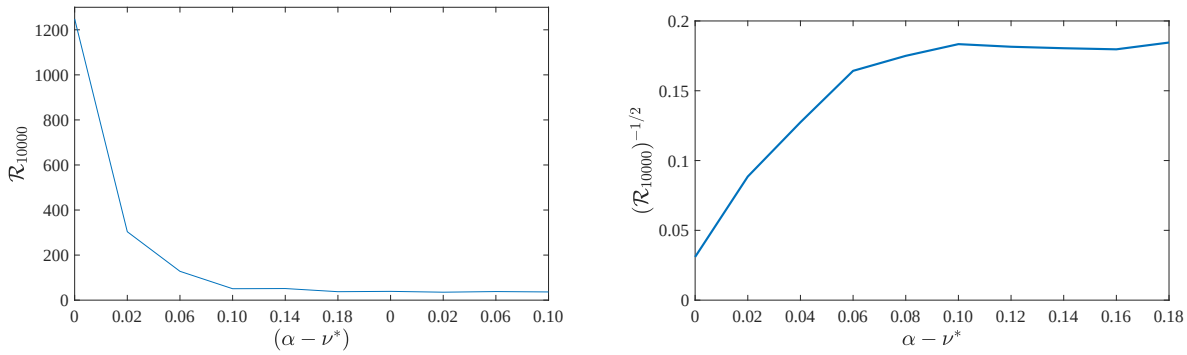

Figure 10: The mean of regret and mean of $1/\sqrt{\mathcal{R}_t}$ over 200 trials as α is varied, plotted against the gap of the optimal arm from the boundary, $\alpha - \nu^*$.

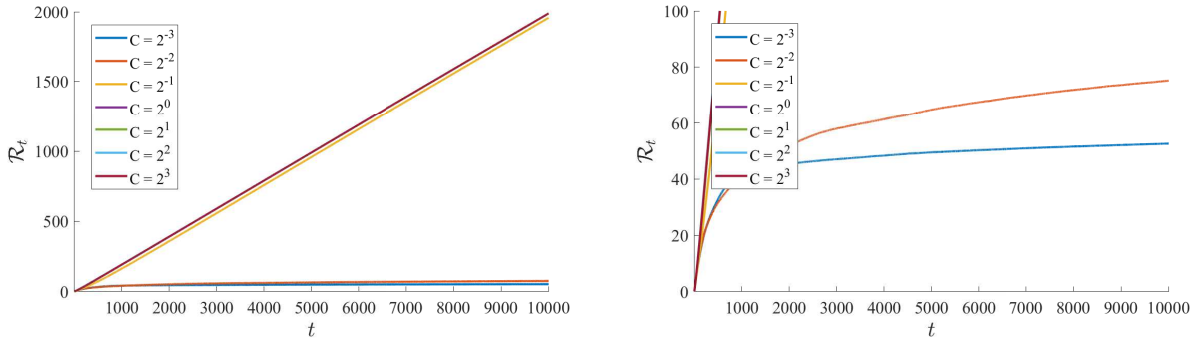

Figure 11: Regret performance as the slack factor C is varied. Right zooms into the bottom half of the left plot. Merans over 200 trials are reported.

One natural approach to address this obstacle is to allow a slack in the safety criterion for TS - we may sample θ_t^k according to the safety posterior, and then instantiate $\Pi_t = \{k : \theta_t^k \leq \alpha + \varepsilon_t^k\}$, where ε_t^k serves as a slack. This raises a design question of how to choose this slack. We empirically investigate the choice of slack $C\text{Dev}_t^k\sqrt{\log t}$, where Dev_t^k is the standard deviation of the safety posterior or arm k at time t. This choice is natural, since this variance determines the scale of fluctuations of the score itself. Figure 11 shows the behaviour obtained as we set $C = 2^i$ for $i \in [-3 : 3]$ for the same data as before, but now with fixed $\alpha = 0.5$.

This plot, while very preliminary, shows an interesting effect in that values of $C \geq 1/2$ again result in large, linear regret. Recall that $C = 0$, which corresponds to no slack, also gives linear regret. It is unclear how robust this effect is, but if true, this observation suggests that tuning this C properly is a subtle problem, and the behaviour is quite sensitive to it, which raises an interesting challenge for further work.