Acting Optimistically in Choosing Safe Actions

Tianrui Chen [1] **Aditya Gangrade** [2] **Venkatesh Saligrama** [1]

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

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

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. 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 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. We also complement the above theoretical study with simulations.

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.

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. 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 spend-

[1] Boston University [2] Carnegie Mellon University. Correspondence to: Aditya Gangrade <agangra2@andrew.cmu.edu>.

ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

ing 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.

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.

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

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

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* $A_t \in [1:K]$. 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 *competitor*, representing the *best safe arm* given the safety constraint and the mean vectors, is defined as

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

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

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

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,

$$N_t^k := \sum_{s<t} \mathbb{1}\{A_t = k\},$$

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

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 introduce the notation

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

where $d(a\|b)$ denotes the KL divergence between Bernoulli laws with means a and b.

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 broad scheme is described in Algorithm 1. 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.

The KL-UCB type bounds take the following form

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

where γ_t trades-off the width and consistency of U, L. These

Algorithm 1 Doubly Optimistic Confidence Bounds

```
Input: $K$, functions $U, L$.
Initialise: $\mathscr{H}_0 \leftarrow \varnothing$
for $t = 1, 2, \dots$ do
    if $t \le K$ then
        $A_t \leftarrow t$
    else
        $\forall k, L_t^k \leftarrow L(t, \mathscr{H}_{t-1}, k)$.
        $\Pi_t \leftarrow \{k : L_t^k \le \alpha\}$.
        $\forall k \in \Pi_t, U_t^k \leftarrow U(t, \mathscr{H}_{t-1}, k)$.
        $A_t \leftarrow \arg\max_{k \in \Pi_t} U_t^k$.
    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
```

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.

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.

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

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

4. Bayesian Methods

This section explores the use of Bayesian methods for safe bandits. It is natural to consider maintaining frequentist and Bayesian indices for the reward and safety-risk, and we present the Thompson Sampling with BayesUCB scheme as an example. 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 $S \sim \mathrm{Bern}(\nu^{A_t})$.

4.1. Thompson Sampling with BAYESUCB

We take the tack of using a *Bayesian confidence bound*, essentially exploiting the BAYESUCB method of Kaufmann et al. (2012). The idea is to choose a δ_t^kth quantile of the posterior $P_{t,\nu}^k$ as a score, where δ_t^k is a schedule that decays with t. This is able to exploit the potentially improved adaptivity of the posterior, but due to δ_t^k 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_t^k, 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 2, where $Q(P, \delta)$ denotes the δth quantile of the law P. We introduce a slight bias in the same for technical convenience.

Algorithm 2 Thompson Sampling with BAYESUCB (TSBU) for Bernoulli Bandits

```
Input: $K$, schedule $\delta_t^k$.
Initialise: $\mathscr{H}_0 \leftarrow \varnothing$.
for $t = 1, 2, \dots$ do
    $\forall k$
    if $S_t^k = 0$ then
        $L_t^k \leftarrow 0$
    else
        $L_t^k \leftarrow Q(\mathrm{Beta}(S_t^k, N_t^k - S_t^k + 1), \delta_t^k)$.
    end if
    $\Pi_t \leftarrow \{k : L_t^k \le \alpha\}$.
    $\forall k \in \Pi_t$, sample $\rho_t^k \sim \mathrm{Beta}(R_t^k + 1, N_t^k - R_t^k + 1)$
    $A_t \leftarrow \arg\max_{k \in \Pi_t} \rho_t^k$.
    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
```

Theorem 3. *For Bernoulli bandits, Algorithm 2, instantiated with $\delta_t^k = (\sqrt{8N_t^k t \log^3 t})^{-1}$ attains the following regret bound for any $\varepsilon > 0$ and any T :*

$$\mathbb{E}[\mathcal{R}_T] \le \sum_{k \ne k^*} \frac{(1+\varepsilon)(\Delta^k \vee \Gamma^k) \log T}{d_<(\mu^k \| \mu^*) \vee {}^2\!/_3 \cdot d_>(\nu^k \| \alpha)} + \xi_k,$$

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

5. Lower Bound

Proposition 4. *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*

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

Since mean regret can be expressed in terms of $\mathbb{E}[N_T^k]$, this also lower bounds regret.

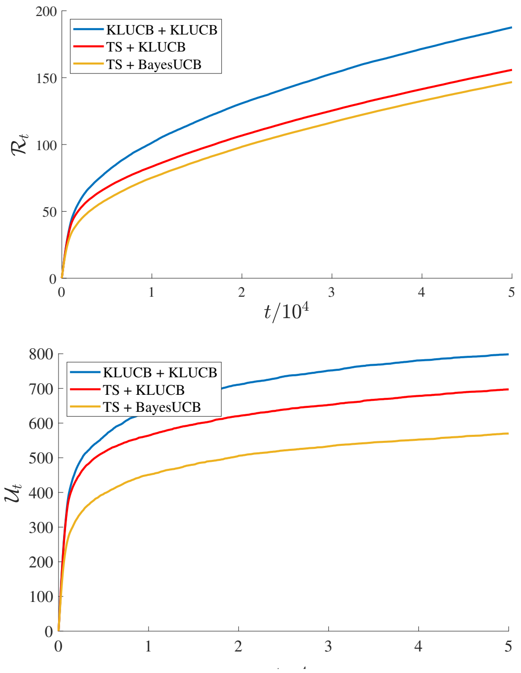

Figure 1. Empirical means over 500 trials of $\mathcal{R}_t$ (top) and $\mathcal{U}_t$ (bottom) for the drug trial data with $\alpha = 0.21$.

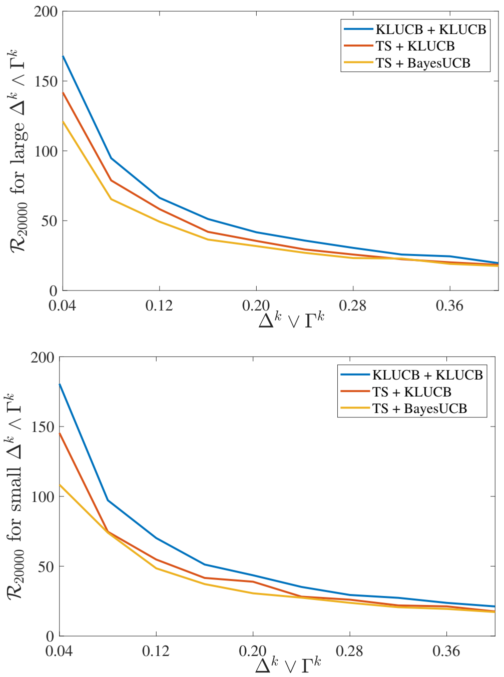

Figure 2. Behaviour of Regret at $T = 20000$ with respect to the maximum gap. Medians over 100 runs are reported.

6. Simulations

6.1. 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.

Observations of Performance From Fig.1, 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 algorithms 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. 2, 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.

Inverse Dependence on Gaps Next, we investigate the dependence of regret on the gaps $\Delta^k \vee \Gamma^k$. First, we will demonstrate that the regret varies with $(\Delta^k \vee \Gamma^k)$ 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. 2 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. 2 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 $t > 10^4$.

Acknowledgements

This research was supported by the Army Research Office Grant W911NF2110246, the National Science Foundation grants CCF-2007350 and CCF-1955981, and the Hariri Data Science Faculty Fellowship Grants, and a gift from the ARM corporation.

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.

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

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.

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

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.

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.

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

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

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

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.