Satisficing Regret Minimization in Bandits: Constant Bound and Light-Tailed Distribution

Qing Feng
School of Operations Research and Information Engineering, Cornell University
qf48@cornell.edu

Tianyi Ma
School of Operations Research and Information Engineering, Cornell University
tm693@cornell.edu

Ruihao Zhu
SC Johnson College of Business, Cornell University
ruihao.zhu@cornell.edu

Motivated by the concept of *satisficing* in decision-making, we consider the problem of satisficing regret minimization in bandit optimization. In this setting, the learner aims at selecting satisficing arms (arms with mean reward exceeding a certain threshold value) as frequently as possible. The performance is measured by *satisficing regret*, which is the cumulative deficit of the chosen arm's mean reward compared to the threshold. We propose `SELECT`, a general algorithmic template for Satisficing REgret Minimization via SampLing and LowEr Confidence bound Testing, that attains constant expected satisficing regret for a wide variety of bandit optimization problems in the *realizable* case (*i.e.*, a satisficing arm exists). Specifically, given a class of bandit optimization problems and a corresponding learning oracle with sub-linear expected (standard) regret upper bound, `SELECT` iteratively makes use of the oracle to identify a potential satisficing arm with low regret. Then, it collects data samples from this arm, and continuously compares the LCB of the identified arm's mean reward against the threshold value to determine if it is a satisficing arm. As a complement, `SELECT` also enjoys the same (standard) regret guarantee as the oracle in the non-realizable case. To further ensure stability of the algorithm, we introduce `SELECT-LITE` that achieves a light-tailed satisficing regret distribution plus a constant expected satisficing regret in the realizable case and a sub-linear expected (standard) regret in the non-realizable case. Notably, `SELECT-LITE` can operate on learning oracles with heavy-tailed (standard) regret distribution. More importantly, our results reveal the surprising compatibility between constant expected satisficing regret and light-tailed satisficing regret distribution, which is in sharp contrast to the case of (standard) regret. Finally, we conduct numerical experiments to validate the performance of `SELECT` and `SELECT-LITE` on both synthetic datasets and a real-world dynamic pricing case study.

Key words: online learning, satisficing, regret, tail risk control

1. Introduction

Multi-armed bandit (MAB) is a classic framework for decision-making under uncertainty. In MAB, a learner is given a set of arms with initially unknown mean rewards. She repeatedly picks an arm to collect its noisy reward and her goal is to maximize the cumulative rewards. The performance is usually measured by the notion of *regret*, which is the difference between the maximum possible

total mean rewards and the total mean rewards collected by the learner. With its elegant form, MAB has found numerous applications in recommendation systems, ads display, and beyond, and many (near-)optimal algorithms have been developed for various classes bandit optimization problems (Lattimore and Szepesvári 2020).

While making the optimal decision is desired in certain high-stakes situations, it is perhaps not surprising that what we often need is just a good enough or *satisficing* alternative (Lufkin 2021). Coined by Nobel laureate Herbert Simon (Simon 1956), the term "satisficing" refers to the decision-making strategy that settles for an acceptable solution rather than exhaustively hunting for the best possible.

The concept of satisficing finds its place in many real world decision-making under uncertainty settings (Caplin et al. 2011) as shown below.

- **Cloud Service:** Consider an operator tuning the configuration of a user-facing cloud service. In each period, the operator selects a configuration, such as a resource allocation, caching rule, batching level, routing policy, or timeout setting, deploys it for a short evaluation window, and observes a noisy measure of service quality, for example the fraction of requests that are completed correctly within a prescribed latency bound. The operator then uses this feedback to choose the configuration in the next period. In this setting, the operator typically does not need to identify the configuration with the highest possible service quality. Instead, the goal is to meet a prespecified service-level objective, such as ensuring that at least 99% of requests satisfy the latency requirement (see, *e.g.*, Chapter 4 of Sloss 2016 and Google 2026). Configurations that reliably meet this target are acceptable, whereas configurations that fall short expose users to degraded service and should be avoided. Therefore, in this setting, the objective is to hit the configurations that meet the reliability requirements as frequently as possible rather than searching for the configuration with the maximum service level.
- **Inventory Management:** Studies show that product stockouts can result in customer disengagement and impair a firm's profitability (Fitzsimons 2000, Jing and Lewis 2011). As such, retailers usually replenish their products periodically to improve service level (*i.e.*, probability of no stockout for any product). Notably, due to the complexity of inventory decisions as well as various constraints (*e.g.*, capacity and inventory cost), it is often very challenging and costly to identify the optimal inventory decision that maximizes service level. Therefore, instead of blindly maximizing service level, retailers typically set a target service level, *e.g.*, the ideal service level in retailing is $90\% \sim 95\%$ (Levy et al. 2023). Consequently, in face of the unknown demand distributions, one can cast the problem into a MAB setting with a *different* objective. In this setting, a retailer starts without complete knowledge of the demand distributions and proceeds by trying various stock level combinations (*i.e.*, arms) of the products over time. Here, the objective

is to hit the target service level as frequently as possible rather than maximizing the service level every time.

- **Product Design:** In fashion product design, designers usually need to design a sequence of products for a group of customers with initially unknown preferences. Since the design of a product is often a highly complex decision involving deciding a combination of many features of a product, identifying the optimal design often involves extensive exploration over many feature combinations, which could be both costly and time-consuming for designers. Therefore, instead of demanding for the best design every time, designers only need to pick a design that can hopefully meet the expectation of a certain portion of the customers for most of the time (Whitenton 2014, Mabey 2022).

The above examples give rise to the study of satisficing bandits (Tamatsukuri and Takahashi 2019, Michel et al. 2023) where the objective is to hit a satisficing arm (*i.e.*, an arm whose mean reward exceeds a certain threshold value) as frequently as possible. The corresponding performance measure is *satisficing regret*, which is the cumulative deficit of the chosen arm's mean reward compared to the threshold. In particular, Michel et al. (2023) develops SAT-UCB[1], an algorithm for finite-armed satisficing bandits: In the *realizable case* (*i.e.*, a satisficing arm exists), the algorithm attains a constant expected satisficing regret, *i.e.*, independent of the length of the time horizon; Otherwise, in the non-realizable case (*i.e.*, the threshold value is out of reach), it preserves the optimal logarithmic expected (standard) regret.

Despite these, existing algorithms usually impose that there is a clear separation between the threshold value and the largest mean reward among all *non-satisficing arms*, which is referred to as the *satisficing gap*. Moreover, the satisficing regret bounds usually scale *inverse proportionally* to the satisficing gap. On the other hand, it is evident that in many bandit optimization problem classes (*e.g.*, Lipschitz bandits and concave bandits), the satisficing gap can simply be 0, making the bounds vacuous. Therefore, it is not immediately clear if SAT-UCB (Michel et al. 2023) or other prior results can go beyond finite-armed bandits and if constant satisficing regret is possible for general classes of bandit optimization problems.

1.1. Main Contributions

In this work, we propose `SELECT`, a novel algorithmic template for Satisficing REgret Minimization via SampLing and LowEr Confidence Bound Testing. More specifically:

- We describe the design of `SELECT` in Section 3. For a given bandit optimization problem class and an associated learning oracle with sub-linear (but not necessarily optimal) expected standard regret guarantee, `SELECT` runs in rounds with the help of the oracle. At the beginning of a round,

[1] Algorithm 2, Section 3.1 in Michel et al. (2023)

SELECT first applies the oracle for a number of time steps and randomly samples an arm from its trajectory as a candidate satisficing arm. Then, it conducts forced sampling by pulling this arm for a pre-specified number of times. Finally, it continues to pull this arm and starts to monitor the resulted lower confidence bound (LCB) of its mean reward. SELECT terminates the current round and starts the next once the LCB falls below the threshold value;

- In Section 4, we establish that in the realizable case, *i.e.*, when a satisficing arm exists, SELECT is able to achieve a constant expected satisficing regret. Notably, the expected satisficing regret bound has no dependence on the satisficing gap (see the forthcoming Remark 4 for a detailed discussion). Instead, it scales inverse proportionally to the *exceeding gap*, which measures the difference between the optimal reward and the threshold value, and is positive in general; Otherwise (*i.e.*, when no satisficing arm exists), SELECT enjoys the same expected standard regret bound as the learning oracle.
- In Section 5, we instantiate SELECT to finite-armed bandits, concave bandits and Lipschitz bandits, and demonstrate the corresponding satisficing and standard regret bounds. We also establish lower bounds of the expected satisficing regret for finite-arm bandits, concave bandits and Lipschitz bandits, showing that the expected satisficing regret bounds established by SELECT are near-optimal for all three bandit problem classes mentioned above;
- While SELECT attains constant satisficing regret in expectation, its performance in individual realizations may not be stable, *i.e.*, the probability of incurring a large satisficing regret is not sufficiently small. To address this, in Section 6, we consider the tail risk control in satisficing regret minimization. To confirm the necessity of new algorithms for light-tailed satisficing regret[2], we first prove that the satisficing regret distribution of SELECT can be heavy-tailed even if we use a learning oracle with light-tailed standard regret distribution. Then, we introduce SELECT-LITE, a modified version of SELECT to obtain a light-tailed satisficing regret distribution. Specifically, with an arbitrary learning oracle with sub-linear expected standard regret bound, SELECT-LITE enjoys the following desirable properties:
  - In the realizable case, SELECT-LITE attains a constant expected satisficing regret bound, and its satisficing regret distribution is light-tailed; and
  - In the non-realizable case, SELECT-LITE still maintains a sub-linear in T standard regret bound.

  Notably, our results do not rely on any assumptions on the tail distribution of the learning oracle's standard regret. In fact, SELECT-LITE is able to make use of a learning oracle with heavy-tailed standard regret to obtain a light-tailed satisficing regret distribution in the realizable case.

[2] The formal definition of light-tailed and heavy-tailed distributions are provided at the beginning of Section 6.

Moreover, in Simchi-Levi et al. (2025), it is shown that any algorithm with a light-tailed standard regret distribution has to suffer a $\Omega(\text{poly}(T))$ instance-dependent expected standard regret[3], rather than a near-optimal $O(\log(T))$ instance-dependent expected standard regret. Because the constant expected satisficing regret bounds are typically instance-dependent, it becomes unclear if a light-tailed satisficing regret distribution will inevitably cause any dependence on T in the expected satisficing regret. Nevertheless, our results show that achieving a light-tailed satisficing regret only worsens the expected satisficing regret by a constant. This thus reveals the compatibility between constant expected satisficing regret and light-tailed satisficing regret distribution. It also indicates a fundamental difference between the notions of standard regret and satisficing regret.

- In Section 7, we conduct numerical experiments to demonstrate the performance of `SELECT` on synthetic instances of finite-armed bandits, concave bandits, and Lipschitz bandits. We use SAT-UCB, SAT-UCB+ (a heuristic with no regret guarantee proposed by Michel et al. 2023), and the respective learning oracle as the benchmarks. Our results reveal that in the realizable case, `SELECT` does achieve constant expected satisficing regret. Moreover, both its expected satisficing regrets (in the realizable cases) and standard regrets (in the non-realizable cases) either significantly outperform the benchmarks' or are close to the best-performing ones'. We also test the performance of `SELECT-LITE` and compare it with `SELECT`. We show that `SELECT-LITE` has better performance in terms of tail distribution control compared to `SELECT` while still maintaining a constant expected satisficing regret comparable to `SELECT`. This indicates that `SELECT-LITE` can provide better protection against tail risks without incurring a significantly larger satisficing regret in expectation.
- To test the performance of our algorithm in real-world instances, in Section 8, we also conduct numerical experiments using a case study of dynamic pricing calibrated from a real-world dataset. Using a real-world avocado sales dataset Kiggins (2018), we show that `SELECT` attains a constant expected satisficing regret in the realizable case. Furthermore, our results reveal that `SELECT` outperforms all other benchmarks both in the realizable and non-realizable cases.
- A preliminary version of this paper (Anonymous 2025) has appeared in the Proceedings of the 13th International Conference on Learning Representations (ICLR 2025). The current paper additionally provides three significant contributions.
  - —First, we go beyond minimizing satisficing regret only in expectation and consider the tail risk control in satisficing regret minimization. Specifically, we build on `SELECT` to establish `SELECT-LITE`, which is able to attain a light-tailed satisficing regret distribution (Theorem 6)

[3] Instance-dependent regret means it is analyzed for a fixed problem instance, *e.g.*, with instance-specific parameters such as $\Delta = \min_{A \neq A^*} r(A^*) - r(A)$, and only its dependence on the time horizon is considered.

while still maintaining a constant expected satisficing regret in the realizable case (Theorem 7) for a wide range of bandit problems;

— Next, in Section 7.4, we test the empirical performance of `SELECT-LITE` using an instance of finite-armed bandit. The results show that `SELECT-LITE` is able to maintain a smaller tail probability in satisficing regret compared to `SELECT` while the expected satisficing regret of `SELECT-LITE` is comparable to that of `SELECT`;

— Finally, in Section 8, we test our algorithm using a case study of dynamic pricing calibrated from a real-world avocado sales dataset. We observe that under this instance `SELECT` is able to significantly outperform all other benchmarks, demonstrating a strong empirical performance of our algorithm in practical bandit problems.

1.2. Related Works

Besides Michel et al. (2023), the most relevant works to ours are Bubeck et al. (2013), Garivier et al. (2019), Tamatsukuri and Takahashi (2019), all of which focus on finite-armed bandits. Bubeck et al. (2013) and Garivier et al. (2019) introduce constant regret algorithms with knowledge of the optimal reward. This can be viewed as a special case of satisficing where the threshold value is exactly the optimal mean reward. Hüyük and Tekin (2021) further extend the algorithm in Garivier et al. (2019) to multi-objective multi-armed bandit settings. Tamatsukuri and Takahashi (2019) also establishes constant regret for finite-armed satisficing bandits when there exists exactly one satisficing arm. As a follow-up paper of Michel et al. (2023), Hajiabolhassan and Ortner (2025) considers satisficing in finite-state MDP.

Among others, Russo and Van Roy (2022) is one of the first to introduce the notion of satisficing regret, but they focus on a time-discounted setting. Furthermore, in Russo and Van Roy (2022), the learner is assumed to know the difference between the values of the threshold and the optimal reward.

More broadly, the concept of satisficing has also been studied by Reverdy and Leonard (2014), Reverdy et al. (2017), Abernethy et al. (2016) in bandit learning. Both Reverdy and Leonard (2014) and Abernethy et al. (2016) focus on maximizing the number of times that the selected arm's *realized reward* exceeds a threshold value. This is equivalent to a MAB problem that sets each arm's mean reward as its probability of generating a value that exceeds the threshold. In view of this, the problems investigated in Reverdy and Leonard (2014) and Abernethy et al. (2016) are closer to conventional MABs. They are also different from Reverdy et al. (2017), Michel et al. (2023) and ours, which use the mean reward of the selected arms to compute the satisficing regret. We note that in Reverdy et al. (2017), a more general Bayesian setting is studied. In this setting, the satisficing regret also takes the learner's belief that some arm is satisficing into account.

The notion of satisficing is also studied in the pure exploration setting (Locatelli et al. 2016, Mukherjee et al. 2017, Kano et al. 2019) as *thresholding bandits*. Their objective is to identify *all* arms with mean reward above a certain satisficing level with limited exploration budget. Unlike the satisficing regret we consider, the performance of their algorithm is measured by simple regret, *i.e.*, the expected number of misidentification made by the final output. Note that any pure exploration algorithm (Audibert and Bubeck 2010) is able to identify the optimal arm or a satisficing arm with high probability after a number of steps. However, due to a small but positive error probability, subsequent exploitation of the identified arm will always incur linear satisficing regret, thus a simple explore-then-exploit approach is unable to obtain a constant satisficing regret bound. Similar observations have also been made in Michel et al. (2023).

Another line of research relevant to our work is tail risk control in bandits. The general goal of this line of research is to design bandit algorithms that achieve light-tailed regret distributions besides minimizing expected regret. Some early works studying the tail distribution of bandit algorithms include Audibert et al. (2009) and Salomon and Audibert (2011), who show that for UCB algorithms, the probability of regret greater than $c(\log(T))^p$ (where $p > 1$ is fixed) decays only polynomially in T. Fan and Glynn (2024) are among the first to establish bandit algorithms with a light-tailed (standard) regret distribution. They focus on the finite-armed bandit setting and show that for informationally-optimized bandit algorithms such as the UCB algorithm and Thompson sampling, the probability of incurring a linear in T regret is at least $\Omega(1/T)$. Some follow-up works include Simchi-Levi et al. (2023) and Simchi-Levi et al. (2025). They show that any finite-armed bandit algorithm with a $O(\log(T))$ instance-dependent expected regret bound has a heavy-tailed regret distribution. They also establish an optimal tradeoff between tail risk, worst-case expected regret bound and instance-dependent expected regret bound. In our paper, we consider tail risk control for satisficing regret minimization. We show that unlike tail risk control in standard regret minimization, where $O(\log(T))$ instance-dependent expected regret bound cannot coexist with light-tailed regret distribution, in satisficing regret minimization there exists algorithm that attains a constant expected satisficing regret and a light-tailed satisficing regret distribution.

2. Problem Formulation

In this section, we present the setup of our problem. Consider a class of bandit optimization problems $(\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is the arm set and $\mathcal{R} \subseteq \{r : \mathcal{A} \to [0,1]\}$ is the class of admissible reward functions. The learner is given a set of feasible arms $\mathcal{A}$ and a class of admissible reward functions $\mathcal{R}$, pulling arm $A \in \mathcal{A}$ in a time step brings a random reward with mean $r(A)$. The learner knows the underlying reward function $r(\cdot)$ belongs to the function class $\mathcal{R}$, but the exact $r(\cdot)$ is initially unknown to the learner. In each time step $t = 1, 2, \ldots, T$, the learner pulls an arm $A_t \in \mathcal{A}$, and then observes a noisy reward Y_t supported on $[0,1]$ with mean $\mathbb{E}[Y_t | A_1, \epsilon_1 \ldots, A_{t-1}, \epsilon_{t-1}, A_t] = r(A_t)$.

Remark 1. *The notion of problem class $(\mathcal{A}, \mathcal{R})$ captures most of the bandit optimization problems in the stationary setting. For example, if $\mathcal{A} = [K]$ and $\mathcal{R} = \{r : [K] \to [0,1]\}$, then the problem class $(\mathcal{A}, \mathcal{R})$ corresponds to finite-armed bandits with K arms; if $\mathcal{A}$ is any convex set in $\mathbb{R}^d$ and $\mathcal{R} = \{r : \mathcal{A} \to [0,1], r$ is concave and 1-Lipschitz$\}$, then the problem class $(\mathcal{A}, \mathcal{R})$ corresponds to concave bandits.*

In this work, we define two notions of regret. First, with the presence of satisficing level S, we consider the satisficing regret introduced in Reverdy et al. (2017), Michel et al. (2023), which is the expected cumulative deficit of the chosen arm's mean reward when compared to S, *i.e.*

$$\texttt{Regret}_S = \sum_{t=1}^{T} \max\{S - r(A_t), 0\}.$$

Because it is possible that no arm achieves the satisficing level, we also follow the conventional bandit literature to define the (standard) regret, which measures the expected cumulative difference of between the chosen arm's mean reward and the optimal mean reward. Specifically, let $A^* = \arg\max_{A \in \mathcal{A}} r(A)$, the standard regret is defined as

$$\texttt{Regret} = T \cdot r(A^*) - \sum_{t=1}^{T} r(A_t).$$

We say that the setting is *realizable* if $r(A^*) \geq S$; Otherwise, it is *non-realizable*. We recall that an arm A is satisficing if $r(A) \geq S$; Otherwise, it is non-satisficing.

Remark 2. *In the definition of satisficing regret, we consider the cumulative deficit of mean reward compared to the satisficing threshold because ideally we would like the arm pull in every time step to be satisficing, thus we penalize every arm pull of non-satisficing arms. Even though in some time steps the decision maker may pull an arm with mean reward over the satisficing threshold, the excessive mean reward cannot compensate for the satisficing regret incurred in other time steps. With this definition of satisficing regret, the problem cannot be trivially solved by running a traditional bandit algorithm. Although any bandit algorithm will gradually converge to the optimal arm and achieve a satisficing total reward with long enough time horizon, as time horizon T increase the number of arm pulls of non-satisficing arm will also increase, incurring a satisficing regret scaling in T. In the forthcoming Section 7 and Section 8, we also numerically show that directly running bandit algorithms can incur substantial satisficing regret.*

Existing works (*e.g.*, Michel et al. 2023) are mostly focused on satisficing regret minimization in finite-armed bandit settings, and usually derive expected satisficing regret bound that scales inversely proportional to the *satisficing gap*

$$\Delta_S = \min\{S - r(A) : r(A) < S\}.$$

However, for bandit optimization problems with large and even infinitely many arms, Δ_S can approach 0 quickly. To address this issue, we define the notion of *exceeding gap* that captures the difference between the optimal reward and S in realizable cases, *i.e.*,

$$\Delta_S^* = r(A^*) - S.$$

Remark 3 (**Shortfalls of Existing Algorithms**). *In what follows, we discuss how existing algorithms such as the ones in Garivier et al. (2019) and Michel et al. (2023) are unable to extend to more general bandit settings. At a high level, both the algorithm in Garivier et al. (2019) and the one in Michel et al. (2023) use the following steps: first, maintain and exploit a large set of candidate satisficing arms containing all arms that are potentially satisficing; second, while exploiting the candidate satisficing arm set, remove any arm that no longer meets the criteria of a candidate satisficing arm; finally, when the candidate arm set becomes empty, uniformly explore the entire arm set until the candidate arm set becomes non-empty again. Specifically, in Garivier et al. (2019), the criteria for a candidate satisficing arm is that the arm's UCB should exceed S, while in Michel et al. (2023), the criteria is that the arm's empirical mean reward should exceed S.*

While the aforementioned idea works well for finite-arm bandits, it can easily fail in many other bandit problem classes, particularly those with continuous arm sets (e.g., concave bandits and Lipschitz bandits). For existing algorithms such as Garivier et al. (2019) and Michel et al. (2023), continuous arm sets presents at least two challenges:

- *With a continuous arm set, the number of arms becomes infinite, and even after applying certain discretization, the algorithm still needs to explore many more arms compared to finite-arm bandit settings when doing uniform exploration, and many of those arms could be non-satisficing. For this reason, uniform exploration could also mean a large number of arm pulls among non-satisficing arms, thus incurring significant satisficing regret. Similarly, removing non-satisficing arms could also be costly, as the candidate satisficing arm set could potentially contain a large number of non-satisficing arms and significant regret may incur when removing these non-satisficing arms one by one.*
- *The issue is further worsened by the lack of separation between mean rewards and the satisficing threshold. Specifically, under continuous arm set, mean rewards of non-satisficing arms can be arbitrarily close to the satisficing threshold S. This brings significant challenges to removal of non-satisficing arms. Specifically, consider a non-satisficing arm whose mean reward is $S - \delta$. Under the criteria for candidate satisficing arms in either Garivier et al. (2019) or Michel et al. (2023), removal of this arm is guaranteed only after $\tilde{\Omega}(1/\delta^2)$ pulls of the arm, and a satisficing regret of $\tilde{\Omega}(1/\delta)$ is incurred when removing such a non-satisficing arm. With the arm set being*

continuous, δ could be arbitrarily small, making the satisficing regret incurred when removing even a single non-satisficing arm arbitrarily large already. We would also like to remark that this issue cannot be solved by simply using a stricter criteria for candidate satisficing arms, e.g., requiring that the arm's LCB should exceed S. With a stricter critera for candidate satisficing arms, the candidate arm set can easily become empty, forcing the algorithm to run the already costly uniform exploration over and over again, thus incurring even larger satisficing regret.

For these reasons, existing algorithms in Garivier et al. (2019) and Michel et al. (2023) are unable to attain a constant expected satisficing regret in bandit classes beyond finite-arm bandits. Addressing these challenges requires significant algorithmic departure from existing approaches.

3. SELECT: Satisficing Regret Minimization via Sampling and Lower Confidence Bound Testing

In this section, we present our algorithmic template SELECT for general classes of bandit optimization problems. When the arm set $\mathcal{A}$ is not too large, one can collect enough data samples for every arm to estimate its mean reward and gradually abandon any non-satisficing arms (see, *e.g.*, Garivier et al. 2019, Michel et al. 2023). In general, however, the arm set can be large and may even contain infinitely many arms (*e.g.*, continuous arm set). Therefore, it becomes impossible to identify all non-satisficing arms. We thus take an alternative approach: The algorithm repeatedly locates a potentially satisficing arm with low (satisficing) regret. Then, it tests if this candidate arm is truly a satisficing arm or not. If not, it kicks off the next round of search.

To ensure fast detection of candidate arms with low regret for a given bandit optimization problem class $(\mathcal{A}, \mathcal{R})$, SELECT makes use of a bandit algorithm for (standard) regret minimization on $(\mathcal{A}, \mathcal{R})$. To this end, we assume access to a blackbox learning oracle that achieves sub-linear standard regret for the bandit optimization problem class $(\mathcal{A}, \mathcal{R})$.

Condition 1. *For problem class $(\mathcal{A}, \mathcal{R})$, there exists a sequence of learning algorithms ALG such that for any reward function $r \in \mathcal{R}$ and any time horizon $t \geq 2$, the expected regret of ALG(t) is bounded by*

$$t \cdot r(A^*) - \mathbb{E}\left[\sum_{s=1}^{t} r(A_s)\right] \leq C_1 t^{\alpha} \log(t)^{\beta}.$$

where $C_1 \geq 1, 1/2 \leq \alpha < 1, \beta \geq 0$ are constants only dependent on the problem class and independent from the time horizon t.

We remark that Condition 1 only asks for sub-linearity in standard regret upper bounds rather than optimality. Therefore, it is an extremely mild condition and can be easily satisfied by a wide range of bandit optimization problem classes and the corresponding bandit learning oracles, *e.g.*, finite-armed bandits, linear bandits, concave bandits, Lipschitz bandits, etc. We will demonstrate this in the forthcoming Section 5 and Section I of the Appendix.

With this, for a problem class $(\mathcal{A}, \mathcal{R})$ and its sub-linear regret learning algorithm ALG, SELECT runs in rounds. In round i, it takes the following steps (see Fig. 1 for a illustration):

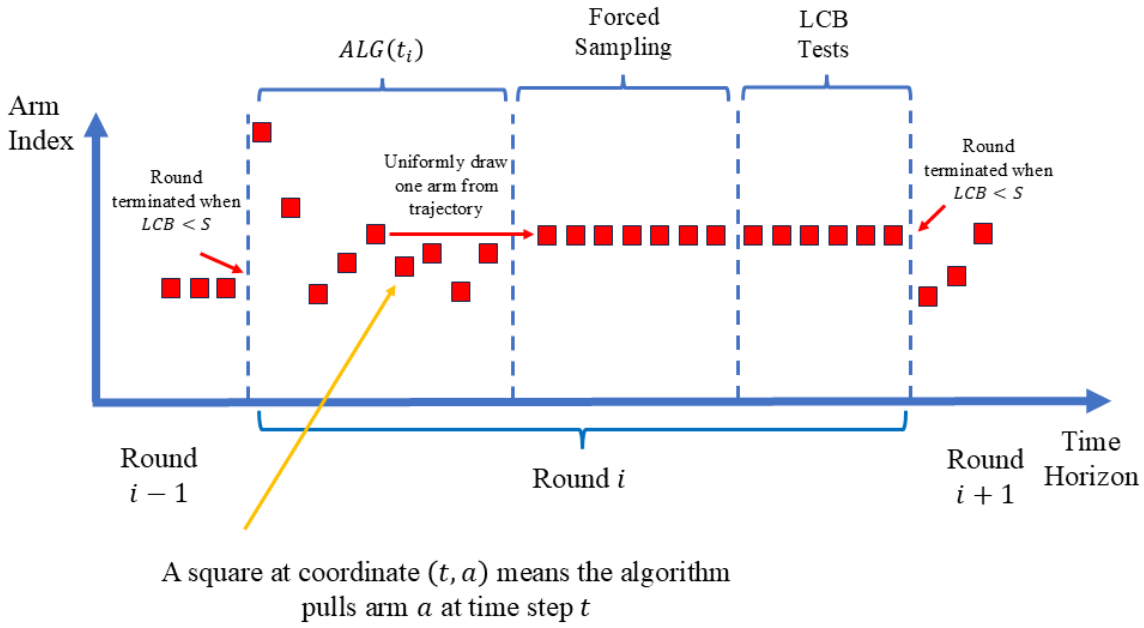


Figure 1 **Illustration of a single round of** SELECT

Step 1. Identify a Potentially Satisficing Arm: Let $\gamma_i = 2^{-i(1-\alpha)/\alpha}$, SELECT follows $\texttt{ALG}(t_i)$ for the first $t_i = \lceil \gamma_i^{-1/(1-\alpha)} \rceil = \lceil 2^{i/\alpha} \rceil$ steps of round i and records its arm selections. Then, it samples an arm $\hat{A}_i$ uniformly random from the trajectory of $\texttt{ALG}(t_i)$ in this round. By virtue of ALG, we find an arm whose mean reward is at most $\tilde{O}(t_i^{\alpha-1}) = \tilde{O}(\gamma_i)$ below $r(A^*)$ in expectation, and only $\tilde{O}(t_i^{\alpha})$ satisficing regret is incurred in this step in the realizable case[4]. We note that in the realizable case, as i increases, $\tilde{O}(\gamma_i)$ will gradually become smaller than Δ_S^*, meaning that the sampled arm $\hat{A}_i$ is more likely to be a satisficing arm and enjoys no satisficing regret;

Step 2. Forced Sampling on the Identified Arm: To validate if $\hat{A}_i$ is a satisficing arm or not, we need to collect enough data from pulling $\hat{A}_i$ before performing any statistical test (see the forthcoming Remark 4 for the reasons). However, when the arm set $\mathcal{A}$ is large, $\texttt{ALG}(t_i)$ might only have pulled $\hat{A}_i$ for limited number of times. To circumvent this, we perform a forced sampling on $\hat{A}_i$ by pulling it for $T_i = \lceil \gamma_i^{-2} \rceil$ times;

Step 3. Lower Confidence Bound Tests: In this step, SELECT kicks off the LCB test as more data is collected from pulling $\hat{A}_i$. We use k (initialized to 0) to denote the number of additional

[4] We adopt the asymptotic notations $O(\cdot), \Omega(\cdot)$, and $\Theta(\cdot)$ as defined in (Cormen et al. 2009). When logarithmic factors are omitted, we write them as $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$, and $\tilde{\Theta}(\cdot)$.

pieces of noisy reward data generated by pulling $\hat{A}_i$ in the current time step and $\hat{r}_i^{\text{tot}}$ to denote the running total reward collected from Step 2 and the current step so far. At the same time, it keeps comparing the LCB of $\hat{A}_i$'s mean reward, *i.e.*, $\hat{r}_i^{\text{tot}}/(T_i+k) - \sqrt{4\log(T_i+k)/(T_i+k)}$, against the threshold value S after acquiring every piece of new data. `SELECT` terminates this round and enters the next whenever the LCB is less than S.

The pseudo-code of `SELECT` is presented in Algorithm 1.

Remark 4 (**Key Novelties of** `SELECT`)**.** *Our algorithm brings together several key ingredients, including sampling from oracle trajectories (Step 1), forced sampling to collect data (Step 2), and the LCB test (Step 3). In what follows, we further comment on the subtleties of each step.*

Step 1. *Prior works on satisficing (Garivier et al. 2019, Michel et al. 2023) mainly use uniform exploration to find candidates of satisficing arms. This approach could incur major satisficing regret when the arm set is large or even possibly infinite, as conducting uniform exploration on a large arm set may result in pulling a large number of non-satisficing arms. To bypass this challenge, in Step 1, we search for candidates using a bandit learning oracle with sub-linear standard regret. Hence, we are able to find candidates of satisficing arms without incurring large regret.*

Step 2. *One challenge from using the LCB test in Step 3 is that it is a more conservative design compared to prior works. In fact, if we directly enter Step 3 without Step 2, the LCB of $r(\hat{A}_i)$ can easily fall below S even if $r(\hat{A}_i) \geq S$. This is because we may not have pulled $\hat{A}_i$ for sufficiently many times in Step 1 and the corresponding confidence interval might be large. As a result, the lower confidence bound of $r(\hat{A}_i)$, which is the difference between empirical mean and confidence radius, can be significantly below S. This indicates, without the forced sampling in Step 2, major satisficing regret can be incurred due to frequent re-start of a new round. With the forced sampled data collected from Step 2,* `SELECT` *ensures that the width of the confidence interval is of order $\tilde{O}(T_i^{-1/2}) = \tilde{O}(\gamma_i)$ (i.e., same as $\mathbb{E}[r(A^*) - r(\hat{A}_i)]$). Consequently, whenever γ_i shrinks to well below Δ_S^*,* `SELECT` *gradually becomes less likely to terminate a round (see the forthcoming Proposition 2);*

Step 3. *In Step 3,* `SELECT` *compares the LCB of $\hat{A}_i$'s mean reward against S to determine if A_i is a satisficing arm. This deviates from prior works that use UCB (Garivier et al. 2019) or empirical mean (Michel et al. 2023). However, it turns out that this is essential in achieving a constant expected satisficing regret that does not scale with $1/\Delta_S$, which can quickly explode in many cases e.g., when the satisficing gaps can be arbitrarly small. As we will see, once entering Step 3,* `SELECT` *can terminate a round within 1 additional time step in expectation when facing a non-satisficing arm (i.e.,* (17) *in the proof of the forthcoming Proposition 1). In contrast, if it follows prior works to use UCB or empirical mean instead, the number of time steps required (and hence, the satisficing regret) will unavoidably scale with $1/\Delta_S$ (see e.g., Theorem 9 of Garivier et al. 2019 or Theorem 1 of Michel et al. 2023).*

Algorithm 1 Satisficing Regret Minimization via Sampling and Lower Confidence Bound Testing (SELECT)

Input: time horizon T, satisficing level S
Set $\gamma_i \leftarrow 2^{-i(1-\alpha)/\alpha}$ for all $i = 1, 2, 3 \ldots$
for round $i = 1, 2, \ldots$ **do**
  Set $t_i \leftarrow \lceil \gamma_i^{-1/(1-\alpha)} \rceil$
  Run ALG(t_i), let $\{\bar{A}_s\}_{s \in [t_i]}$ be the trajectory of arm pulls
  Sample q uniformly at random from $\{1, 2, \ldots, t_i\}$ and set $\hat{A}_i \leftarrow \bar{A}_q$
  Set $T_i \leftarrow \lceil \gamma_i^{-2} \rceil$, $k \leftarrow 0$
  Let t be the current time step, pull arm $\hat{A}_i$ for the next T_i time steps, and set $\hat{r}_i^{\text{tot}} \leftarrow \sum_{s=t}^{t+T_i-1} Y_s$
  while $LCB(\hat{A}_i) \geq S$ **do**
    Set $k \leftarrow k + 1$
    Pull arm $\hat{A}_i$ and observe Y_t, where t is the current time step
    Update $\hat{r}_i^{\text{tot}} \leftarrow \hat{r}_i^{\text{tot}} + Y_t$
    Set

$$LCB(\hat{A}_i) \leftarrow \frac{\hat{r}_i^{\text{tot}}}{T_i + k} - \sqrt{\frac{4 \log(T_i + k)}{T_i + k}}$$

  end while
end for

4. Regret Analysis

In this section, we analyze the regret bounds of SELECT. We begin by providing an upper bound of expected satisficing regret in the realizable case, *i.e.*, $r(A^*) \geq S$.

Theorem 1. *If $r(A^*) \geq S$, then the expected satisficing regret of SELECT is bounded by*

$$\mathbb{E}[\textit{Regret}_S] \leq \min \left\{ C_1^{\frac{1}{1-\alpha}} \left(\frac{1}{\Delta_S^*} \right)^{\frac{\alpha}{1-\alpha}} \cdot \text{polylog} \left(\frac{C_1}{\Delta_S^*} \right), C_1 T^{\alpha} \cdot \text{polylog}(T) \right\}. \tag{1}$$

We remark that, from the first term on the RHS of (1), SELECT can achieve a constant (w.r.t. T) expected satisficing regret in the realizable case. Moreover, even when T is relatively small and the constant expected satisficing regret guarantee is loose compared to the oracle's regret bound, it is able to adapt to the oracle's performance and achieve a sub-linear in T bound on expected satisficing regret. More specifically, when the exceeding gap Δ_S^* is large, *i.e.*, $\Delta_S^* \geq C_1 T^{-(1-\alpha)}$, the regret bound w.r.t. Δ_S^* is the smaller one in the two bounds. Therefore in this case the satisficing regret bound scales polynomially in $1/\Delta_S^*$ and does not increase even when the time horizon T increases. When the exceeding gap Δ_S^* is small, *i.e.*, $\Delta_S^* < C_1 T^{-(1-\alpha)}$, the regret bound w.r.t. T is the smaller one in the two bounds. Therefore in this case the satisficing regret bound scales sublinearly in T and does not increase as Δ_S^* gets smaller.

Proof Sketch. A complete proof for Theorem 1 is provided in Section A. As discussed in Section 3, there are two key intuitions why SELECT attains low expected satisficing regret: first, by virtue of the standard regret minimization algorithm ALG, Step 1 and Step 2 of SELECT does not incur much

satisficing regret, and Step 3 can also quickly remove non-satisficing arms without incurring much satisficing regret; second, as the rounds proceed, `SELECT` becomes more likely to acquire a candidate arm that is truly satisficing, and the candidate arm is also less likely to be removed by the LCB test in Step 3. In what follows, we present two key results that formalize the two intuitions.

To formalize the first intuition, we use the following proposition to establish an upper bound on the expected satisficing regret incurred in each round.

Proposition 1. *If $r(A^*) \geq S$, then the expected satisficing regret incurred in round i is bounded by*

$$\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

The proof of this proposition is provided in Section A.1 of the appendix.

To formalize the second intuition, we use the following proposition to show that once `SELECT` runs for enough rounds so that $\gamma_i = \tilde{O}(\Delta_S^*)$, it becomes less likely to start a new round.

Proposition 2. *If $r(A^*) \geq S$, then for every i that satisfies*

$$\frac{32C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^{\max\{\beta,1/2\}} \leq \Delta_S^*\,, \tag{2}$$

the probability that round i ends within the time horizon T (conditioned on the event that round i is started within the time horizon) is upper bounded by $1/4$.

The proof of this proposition is provided in Section A.2 of the appendix. Since each round of `SELECT` runs independently, Proposition 2 indicates that the total number of rounds for `SELECT` can be upper-bounded by a shifted geometric random variable with success rate $3/4$.

With Proposition 1 and Proposition 2, we are ready to finish the proof of Theorem 1. We denote i_0+1 as the minimum integer such that (2) holds. We prove that $2^{i_0(1-\alpha)/\alpha} = 1/\gamma_{i_0} = C_1/\Delta_S^* \cdot \mathrm{polylog}(C_1/\Delta_S^*)$. We further have $i_0 = \Theta(\log(C_1/\Delta_S^*))$. Therefore by Proposition 1 the expected satisficing regret incurred in the first i_0 rounds is bounded by

$$\sum_{i=1}^{i_0}\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\frac{\alpha}{1-\alpha}}\log(4/\gamma_i)^\beta \leq C_1^{\frac{1}{1-\alpha}}\left(\frac{1}{\Delta_S^*}\right)^{\frac{\alpha}{1-\alpha}}\cdot\mathrm{polylog}\left(\frac{C_1}{\Delta_S^*}\right).$$

Then we bound the expected satisficing regret incurred after round i_0. By Proposition 2, round i_0+k-1 ends within the time horizon and `SELECT` further enters round i_0+k with probability at most 4^{1-k}. Therefore the expected satisficing regret after round i_0 is upper bounded by

$$\sum_{k=1}^{\infty}\frac{8C_1}{(1-\alpha)^\beta}\gamma_{i_0+k}^{-\frac{\alpha}{1-\alpha}}\log(4/\gamma_{i_0+k})^\beta\cdot 4^{1-k} = \frac{32C_1}{(1-\alpha)^\beta}\gamma_{i_0}^{-\frac{\alpha}{1-\alpha}}\sum_{k=1}^{\infty}2^{-k}\left(\frac{(1-\alpha)(i_0+k)}{\alpha}+1\right)^\beta\log(2)^\beta.$$

We have shown that $\gamma_{i_0}^{-\alpha/(1-\alpha)} = (C_1/\Delta_S^*)^{\alpha/(1-\alpha)}\cdot\mathrm{polylog}(C_1/\Delta_S^*)$. We further show that the summation above is polynomial in i_0, thus polynomial in $\log(C_1/\Delta_S^*)$. Therefore the expected satisficing

regret is upper bounded by $C_1^{1/(1-\alpha)}(1/\Delta_S^*)^{\alpha/(1-\alpha)} \cdot \text{polylog}(C_1/\Delta_S^*)$. Therefore the expected satisficing regret is upper bounded by the first term in the regret bound.

We further show that the expected satisficing regret is upper bounded by the second term in the regret bound. Let $i_{\max}$ be the total number of rounds for SELECT. We show that $t_{i_{\max}} = \gamma_{i_{\max}}^{-1/(1-\alpha)} = O(T)$, thus $i_{\max} = O(\log(T))$. By Proposition 1 the expected satisficing regret of each round is upper bounded by $\tilde{O}(C_1\gamma_i^{-\alpha/(1-\alpha)}) = \tilde{O}\left(C_1 t_i^{\alpha}\right) \leq \tilde{O}\left(C_1 T^{\alpha}\right)$. Since the total number of rounds is upper bounded by $O(\log(T))$, the expected satisficing regret of SELECT is also upper bounded by $C_1 T^{\alpha} \cdot \text{polylog}(T)$. □

We also show that, in the non-realizable case, SELECT enjoys the same regret bound as the oracle in Condition 1 (The proof is provided in Section B of the appendix).

Theorem 2. *If $r(A^*) < S$, the standard regret of SELECT is bounded by*

$$\mathbb{E}[\textit{Regret}] \leq C_1 T^{\alpha} \cdot \text{polylog}(T).$$

Proof of Theorem 2 is provided in Section B. Theorem 2 implies that the expected standard regret of SELECT is sub-linear in T in the non-realizable case. The expected standard regret bound of SELECT matches that of the learning oracle within logarithmic factors of T, indicating a similar performance guarantee of SELECT compared to the standard regret minimization algorithm in the corresponding class of bandit optimization problems.

With the above results, we establish that under a very general condition, SELECT does achieve constant expected satisficing regret in the realizable case without compromising the standard regret guarantee in the non-realizable case.

5. Examples

In what follows, we instantiate SELECT to several popular classes of bandit optimization problems, including finite-armed bandits, concave bandits, and Lipschitz bandits. Along the way, we showcase how our results enable constant expected satisficing regret for bandits with large and even infinite arm set, which was not achieved by existing algorithms. We remark that the examples given here are non-exhaustive and similar results can be derived for other classes of bandit optimization problems such as linear bandits. In all three examples of this section, we assume the mean reward of each arm is in $[0, 1]$.

1. Finite-Armed Bandits: Consider any instance of finite-armed bandit with K arms, *i.e.*, $\mathcal{A} = [K]$. In this case, both the UCB algorithm (see, *e.g.*, Bubeck and Cesa-Bianchi 2012) and Thompson sampling (see, *e.g.*, Agrawal and Goyal 2017) achieve a regret bound of $O(\sqrt{KT\log(T)})$. Combining them with Theorems 1 and 2, we have the following corollary. At the end of this section, we also provide some discussions on the lower bounds of the expcted satisficing regret.

Corollary 1. *By using either the UCB algorithm or Thompson sampling as* `ALG`*, if* $r(A^*) \geq S$*, the expected satisficing regret of* `SELECT` *is bounded by*

$$\mathbb{E}[\texttt{Regret}_S] \leq \min\left\{\frac{K}{\Delta_S^*}\cdot \text{polylog}\left(\frac{K}{\Delta_S^*}\right), \sqrt{KT}\cdot\text{polylog}(T)\right\};$$

If $r(A^*) < S$*, the standard regret of* `SELECT` *is bounded by* $\mathbb{E}[\texttt{Regret}] \leq \sqrt{KT}\cdot\text{polylog}(T)$.

Remark 5 (**Comparison with Garivier et al. 2019, Michel et al. 2023**)**.** *Garivier et al. (2019), Michel et al. (2023) also provide algorithms for the expected satisficing regret in finite-armed bandit settings. While our expected satisficing regret bound is incomparable to Garivier et al. (2019), which attains a* $O(K/\Delta_S)$ *expected satisficing regret, we provide major improvement over Michel et al. (2023), which achieves* $O(K/\Delta_S + K/(\Delta_S^*)^2)$ *expected satisficing regret. Compared to Michel et al. (2023), our regret bound remove the dependence on* Δ_S *and the additional* $1/\Delta_S^*$ *factor. We also point out that if we want to achieve the same expected satisficing regret as Garivier et al. (2019), we can simply change the LCB test in Step 3 to a UCB test.*

Remark 6 (**Best-of-Both-Worlds**)**.** *When comparing the our expected regret bound in Corollary 1 with the one in Garivier et al. (2019), one can see that our algorithm* `SELECT` *potentially has a better performance when* Δ_S^* *is large and* Δ_S *is small, while the algorithm in Garivier et al. (2019) potentially has a better performance when* Δ_S^* *is small and* Δ_S *is large. In order to establish an algorithm that attains a "best of both worlds" and works well in both cases, in Section N of the appendix, we propose* `SELECT-BoBW`*, an adapted version of* `SELECT` *for finite-arm bandits that attains an expected satisficing regret bound* $O(\min\{K/\Delta_S, K/\Delta_S^*\})$*. , a regret bound that is the Best-of-Both-Worlds, improving upon both Garivier et al. (2019), Michel et al. (2023). Specifically,* `SELECT-BoBW` *is modified based on* `SELECT` *by adding a UCB test in the forced sampling in Step 2, which compares the UCB of candidate arm mean reward against the threshold* S*, during forced sampling. Whenever the UCB of the candidate satisficing arm falls below* S*, the forced sampling is terminated early and* `SELECT-BoBW` *immediately proceeds to the next round. We present a full discussion of* `SELECT-BoBW` *in Section N in the appendix.*

2. Concave Bandits: Consider any instance of the concave bandit, *i.e.*, $\mathcal{A} \subset \mathbb{R}^d$ is a bounded convex set, and $r(A)$ is a concave and 1-Lipschitz continuous function defined on $\mathcal{A}$. Agarwal et al. (2011) gives an algorithm that enjoys $\text{poly}(d)\sqrt{T}\cdot\text{polylog}(T)$ regret. Together with Theorems 1 and 2, we have the following corollary.

Corollary 2. *By using the algorithm in Agarwal et al. (2011) as* `ALG`*, if* $r(A^*) \geq S$*, then the expected satisficing regret of* `SELECT` *is bounded by*

$$\mathbb{E}[\texttt{Regret}_S] \leq \min\left\{\frac{\text{poly}(d)}{\Delta_S^*}\cdot \text{polylog}\left(\frac{d}{\Delta_S^*}\right), \text{poly}(d)\sqrt{T}\cdot\text{polylog}(T)\right\};$$

If $r(A^*) < S$*, the standard regret of* `SELECT` *is bounded by* $\mathbb{E}[\texttt{Regret}] \leq \text{poly}(d)\sqrt{T}\cdot\text{polylog}(T)$.

3. Lipschitz Bandits: Consider any instance of the Lipschitz bandit in d dimensions, *i.e.*, $\mathcal{A}=[0,1]^d$, and $r(A)$ is an L-Lipschitz function (here Lipschitz functions are defined in the sense of ∞-norm). Bubeck et al. (2011) introduces a uniformly discretized UCB algorithm to achieve a $O(L^{d/(d+2)}T^{(d+1)/(d+2)}\sqrt{\log(T)})$ regret upper bound. Together with Theorem 1 and Theorem 2, we have the following corollary.

Corollary 3. *By using the UCB algorithm with uniform discretization in Bubeck et al. (2011) as* `ALG`*, if $r(A^*)\geq S$, then the expected satisficing regret of* `SELECT` *is bounded by*

$$\mathbb{E}[\texttt{Regret}_S]\leq \min\left\{\frac{L^d}{(\Delta_S^*/2)^{d+1}}\cdot \text{polylog}\left(\frac{L}{\Delta_S^*}\right), L^{\frac{d}{d+2}}T^{\frac{d+1}{d+2}}\cdot \text{polylog}(T)\right\};$$

If $r(A^*)<S$, the standard regret of `SELECT` *is bounded by $\mathbb{E}[\texttt{Regret}]\leq L^{\frac{d}{d+2}}T^{\frac{d+1}{d+2}}\cdot\text{polylog}(T)$.*

Remark 7. *Recall that $\Delta_S=\min\{S-r(A):r(A)<S\}$, one can easily verify that in the realizable case, $\Delta_S=0$ for both concave bandits and the Lipschitz bandits. As such, one cannot directly apply the results in Garivier et al. (2019), Michel et al. (2023) to acquire a constant expected satisficing regret. With the notion of exceeding gap Δ_S^* and* `SELECT`*, we establish constant expected satisficing regret bounds for these two classes of bandit optimization problems.*

5.1. Lower Bounds for Satisficing Regret Minimization

To complement our main results, we also present the expected satisficing regret lower bounds for finite-arm bandits, bandits with concave rewards and Lipschitz bandits. By establishing lower bounds on the expected satsificing regret, we show that `SELECT` attains a near-optimal expected satisficing regret bound for all three aforementioned bandit problem classes.

The first one is an expected satisficing regret lower bound for the finite-armed bandits.

Theorem 3 (Finite-Armed Bandits)**.** *For every non-anticipatory learning algorithm π, every $0<\Delta<1/20$, every $K\in\mathbb{Z}_+$, $K\geq 2$, and $T\geq K/\Delta^2$, there exists an instance of finite-arm bandit with K arms such that $r(A^*)-S=\Delta$, and the expected satisficing regret incurred by π is at least $\Omega(K/\Delta)$.*

Theorem 3 establishes a $\Omega(K/\Delta_S^*)$ lower bound on the expected satisficing regret for finite-arm bandits. The lower bound established in Theorem 3 matches the expected satisficing regret upper bound of `SELECT` in Corollary 1 up to a logarithmic factor of K/Δ_S^*. This also suggests that the expected satisficing regret bound of `SELECT` is near-optimal up to logarithmic factors of K/Δ_S^*.

Remark 8. *In Michel et al. (2023), the authors adapt the results from Bubeck et al. (2013) (for MAB with known optimal mean reward) to establish a $\Omega(K/\Delta_S)$ expected satisficing regret lower bound for finite-armed bandits. This is different than the one in Theorem3, which is $\Omega(K/\Delta)$ with Δ being the exceeding gap (i.e., Δ_S^*). This difference originates from the lower bound instances. The lower bound instance in Bubeck et al. (2013) (when adapted to satisficing bandits) sets the threshold*

value to the optimal mean reward, i.e., $S = r(A^*)$, which makes both the expected satisficing regret and standard regret lower bounds to be $\Omega(K/\Delta_S)$. In our lower bound proof, we allow S to be smaller than $r(A^*)$, which leads to a lower bound that depends on the difference between these two quantities.

Next, we provide an expected satisficing regret lower bound for bandits with concave rewards.

Theorem 4 (Bandits with Concave Rewards)**.** *For every non-anticipatory learning algorithm π, every $0 < \Delta < 1/162$ and every $T \geq 1/\Delta^2$, there exists an instance of one-dimensional bandit with concave reward such that $r(A^*) - S = \Delta$, and the expected satisficing regret incurred by π is at least $\Omega(1/\Delta)$.*

Theorem 4 establishes a $\Omega(1/\Delta_S^*)$ lower bound on the expected satisficing regret for concave bandits. The lower bound established in Theorem 4 matches the expected satisficing regret upper bound of `SELECT` in Corollary 2 up to a polynomial factor of d and a logarithmic factor of d/Δ_S^*. If we consider the dimension d as a constant, then the expected satisficing regret bound of `SELECT` for concave bandits is near-optimal up to a logarithmic factor of d/Δ_S^*.

Finally, we provide an expected satisficing regret lower bound for Lipschitz bandits.

Theorem 5 (Lipschitz Bandits)**.** *For every non-anticipatory learning algorithm π, every $0 < \Delta < 1/20$, every $d \in \mathbb{Z}_+$, Lipschitz constant $L \geq 1$ and time horizon $T \geq L^d/\Delta^{d+2}$, there exists an instance of d-dimensonal Lipschitz bandit such that $r(A^*) - S = \Delta$, and the expected satisficing regret incurred by π is at least $\Omega(L^d/\Delta^{d+1})$.*

Theorem 5 establishes a $\Omega(L^d/(\Delta_S^*)^{d+1})$ lower bound on the expected satisficing regret for Lipschitz bandits. The lower bound established in Theorem 5 matches the expected satisficing regret upper bound of `SELECT` in Corollary 3 up to a logarithmic factor of L/Δ_S^*. This also suggests that the expected satisficing regret upper bound provided by `SELECT` is near-optimal up to a logarithmic factor of L/Δ_S^*.

6. Tail Risk Control in Satisficing Regret Minimization

While `SELECT` achieves a constant satisficing regret in expectation, it is unclear if it enjoys a light-tailed distribution on the satisficing regret (i.e., the realized satisficing regret is unlikely to be much larger than the expectation), a highly desirable property in maintaining the stability of an algorithm (Fan and Glynn 2024, Simchi-Levi et al. 2023). We say the satisficing regret of an algorithm is *light-tailed* if the probability of satisficing regret being greater than x decays at a rate of $\exp(-x^\zeta)$ for some $\zeta > 0$ as x increases. In other words, the satisficing regret distribution is light-tailed if there exists $\zeta \in (0,1)$ and Λ_1, Λ_2, such that for any T and any $0 \leq x \leq T$,

$$\Pr(\texttt{Regret}_S > x) \leq \Lambda_1 \exp\left(-x^\zeta/\Lambda_2\right). \tag{3}$$

Here $\Lambda_1, \Lambda_2 > 0$ are parameters independent of x. We remark that Λ_1 and Λ_2 can depend on the length of the time horizon T, but we require that Λ_1 could scale at most polynomially with T,

and Λ_2 could scale at most poly-logarithmically with T. On the other hand, any algorithm whose satisficing regret distribution does not satisfy (3) is said to have a *heavy-tailed* satsficing regret distribution

A light-tailed satisficing regret distribution is critical in many practical settings of satisficing regret minimization, particularly in applications sensitive to tail risk such as healthcare and supply chain. In these applications, a light-tailed satisficing regret distribution reduces the possibility of suffering substantial losses for the decision maker, thus providing the decision maker with more protection against extreme scenarios. For instance, in the inventory management example discussed in Section 1, if an algorithm is able to achieve a light-tailed satisficing regret distribution, it indicates it is highly unlikely for the retailers to frequently fall short of the designated service level over time.

6.1. Key Challenges and Insufficiency of `SELECT`

Based on prior results (see, *e.g.*, Fan and Glynn 2024, Simchi-Levi et al. 2023, Simchi-Levi et al. 2025), achieving a light-tailed satisficing regret distribution while still maintaining a constant expected satisficing regret can be extremely challenging due to the following reasons:

1. **Tradeoff between Instance-Dependency and Tail Risk:** Existing works show that in standard regret minimization for finite-armed bandits, a worse instance-dependent expected standard regret is inevitable for any algorithm with a light-tailed standard regret distribution. In particular, Simchi-Levi et al. (2025) prove that any algorithm with a light-tailed standard regret distribution has to suffer a $\Omega(\text{poly}(T))$ instance-dependent expected standard regret, rather than a near-optimal $O(\log(T))$ instance-dependent expected standard regret for the UCB algorithm and Thompson sampling. Since existing constant expected satisficing regret bounds (*e.g.*, Garivier et al. 2019, Michel et al. 2023, and ours) are all instance-dependent, one may conjecture that in satisficing regret minimization, a light-tailed satisficing regret distribution could also inevitably come at a cost of worse (*e.g.*, polylog(T) or poly(T)) instance-dependent expected satisficing regret. Then, it becomes unclear whether a constant expected satisficing regret bound can coexist with a light-tailed satisficing regret distribution even in the finite-armed bandit setting.
2. **Heavy-Tailed Learning Oracle:** To the best of our knowledge, existing works on tail risk control for standard regret focus mostly on finite-armed bandits and linear bandits, and it remains unclear whether there exists learning oracles with light-tailed regret distribution for more general classes of bandit optimization problems. Therefore, we may have to start with learning oracles with potentially heavy-tailed standard regret distribution. This makes our algorithm design even more challenging in achieving a light-tailed satisficing regret distribution for more general classes of bandit optimization problems.

Before proceeding, we first confirm the necessity of new algorithmic techniques in achieving a light-tailed satisficing regret distribution. We use the following proposition to show that the satisficing regret distribution of SELECT can indeed be heavy-tailed even if the learning oracle ALG used is already light-tailed.

Proposition 3. *For every finite-armed bandit learning oracle with $\tilde{O}(\sqrt{T})$ standard regret, if the oracle selects each arm at least once with probability at least $1/2$, then there exists an instance of two-armed bandit such that for all $T \in \mathbb{Z}_+$, the satisficing regret of SELECT satisfies*

$$Pr\left(\texttt{Regret}_S \geq \frac{T}{14}\right) \geq \frac{1}{4}\exp(-2\log(T)^2).$$

The proof of Proposition 3 is provided in Section D of the appendix. Proposition 3 states that the satisficing regret distribution resulted from SELECT is heavy-tailed under mild conditions of the learning oracle. In fact, such conditions in Proposition 3 hold widely in most of the existing algorithms for finite-armed bandits, such as the UCB algorithm, Thompson sampling, and even for the algorithm in Simchi-Levi et al. (2025), whose standard regret distribution is known to be light-tailed. This suggests that simply applying SELECT does not guarantee a light-tailed satisficing regret distribution, even if the learning oracle adopted is light-tailed. Therefore, a new algorithm is needed if we want to ensure a light-tailed satisficing regret distribution.

6.2. SELECT-LITE: SELECT with Light-Tailed Satisficing Regret Distribution

In this section, we present SELECT-LITE, an augmented algorithmic template with a light-tailed satisficing regret distribution for more general classes of bandit optimization problems.

We note that there are two factors that prevent SELECT from achieving a light-tailed satisficing regret distribution. First, the length t_i of Step 1 and the length T_i of Step 2 increase exponentially as rounds proceed. Since larger t_i and T_i often result in larger tail risk, the rapid growth of t_i and T_i can lead to uncontrolled tail risk, eventually resulting in a heavy-tailed satisficing regret distribution. Second, in the LCB test in Step 3 of SELECT, we construct a sequence of confidence intervals for the candidate satisficing arm such that the k-th confidence interval contains the true mean reward with probability $1 - \text{poly}(1/k)$. Although we have shown that these confidence intervals guarantee removal of a non-satisficing candidate arm in constant time steps in expectation, the distribution of the number of steps in the LCB test could also be heavy-tailed, potentially leading to a heavy-tailed satisficing regret distribution.

In observance of these, we make two modifications to SELECT to make the satisficing regret distribution light-tailed. First, in the modified algorithm, we set the length t'_i of Step 1 and the length T'_i of Step 2 to scale polynomially with the number of rounds i instead of exponentially. Second, we use larger confidence intervals in the LCB test such that the k-th confidence interval

contains the mean reward with probability at least $1-\exp(-\Omega(k^\zeta))$. With this configuration of confidence intervals, if the candidate satisficing arm is non-satisficing, then the probability that the LCB test lasts for more than k time steps is at most $\exp(-\Omega(k^\zeta))$, thus we are able to guarantee that the number of steps for the LCB test in each round follows a light-tailed regret distribution.

We refer to this algorithm as `SELECT-LITE`, and the complete algorithm is provided in Algorithm 2. Here, $\Gamma(z)$ is the gamma function given by $\Gamma(z)=\int_0^\infty t^{z-1}\exp(-t)\mathrm{d}t$.

Algorithm 2 Satisficing Regret Minimization with Light-Tailed Satisficing Regret Distribution (`SELECT-LITE`)

Input: time horizon T, satisficing level S, tail distribution parameter ζ
Set $\gamma_i \leftarrow i^{-(1/\zeta-1)(1-\alpha)}$ for all $i=1,2,3\ldots$
for round $i=1,2,\ldots$ **do**
  Set $t_i' \leftarrow \lceil \gamma_i^{-1/(1-\alpha)} \rceil$
  Run $\texttt{ALG}(t_i')$, let $\{\bar{A}_s\}_{s\in[t_i']}$ be the trajectory of arm pulls
  Sample q uniformly at random from $\{1,2,\ldots,t_i'\}$ and set $\hat{A}_i \leftarrow \bar{A}_q$
  Set $T_i' \leftarrow \lceil \gamma_i^{-2} \rceil$, $k \leftarrow 0$
  Let t be the current time step, pull arm $\hat{A}_i$ for the next T_i' time steps, and set $\hat{r}_i^{\text{tot}} \leftarrow \sum_{s=t}^{t+T_i'-1} Y_s$
  while $LCB(\hat{A}_i) \geq S$ **do**
    Set $k \leftarrow k+1$
    Pull arm $\hat{A}_i$ and observe Y_t, where t is the current time step
    Update $\hat{r}_i^{\text{tot}} \leftarrow \hat{r}_i^{\text{tot}} + Y_t$
    Set

$$LCB(\hat{A}_i) \leftarrow \frac{\hat{r}_i^{\text{tot}}}{T_i'+k} - \sqrt{\frac{k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{T_i'+k}}$$

  end while
end for

6.3. Tail Risk and Regret Analysis of `SELECT-LITE`

In this section, we provide a tail risk and regret analysis of `SELECT-LITE`. We show that `SELECT-LITE` has the following desirable properties: in the realizable case, `SELECT-LITE` enjoys both a light-tailed satisficing regret distribution and a constant expected satisficing regret bound at the same time; in the non-realizable case, `SELECT-LITE` still maintains a sub-linear standard regret bound.

We first use the following theorem to show that `SELECT-LITE` indeed has a light-tailed satisficing regret distribution in the realizable case.

Theorem 6. *If $r(A^*) \geq S$, then for all $x>0$, the probability that satisficing regret of `SELECT-LITE` exceeds x is bounded by*

$$Pr(\texttt{Regret}_S > x) \leq \exp\left(2i_0+4-\left(\frac{(x-x_0)_+}{9+(2(1-\zeta)/3)^{-1}(1-\alpha)^{-1}}\right)^\zeta\right) + \exp\left(i_0+2-\frac{(x-x_0)_+^\zeta}{4}\right),$$

where $i_0 = O\left(\left(1/\Delta_S^* \cdot \log(1/\Delta_S^*)^{\beta'}\right)^{\frac{\zeta}{(1-\zeta)^2(1-\alpha)}}\right)$ *and* $x_0 = 6(i_0 + \zeta(i_0+1)^{1/\zeta})$ *are constants independent of* x *and* T.

Proof Sketch. The complete proof of Theorem 6 is provided in Section E of the appendix. Similar to Proposition 2, we first prove that in `SELECT-LITE`, the probability of round i terminating within the time horizon is at most $1/4$ for all large enough i.

Proposition 4. *If* $r(A^*) \geq S$*, then for all* i *that satisfies*

$$\frac{32C_1(1-\zeta)^{-\zeta/2}}{(1-\alpha)^{\beta}}\sqrt{2+2\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}\gamma_i^{1-\zeta}\log(4/\gamma_i)^{\max\{\beta,1/2\}} \leq \Delta_S^*, \tag{4}$$

the probability that round i *ends within the time horizon* T *(conditioned on the event that round* i *is started within the time horizon) is upper bounded by* $1/4$.

Proposition 4 implies that the total number of rounds for `SELECT-LITE` can be upper-bounded by a shifted geometric random variable with success rate $3/4$. Proof of Proposition 4 is provided in Section E.1 of the appendix.

The inequality in the theorem is trivial when $x < x_0$, thus we focus on the case where $x \geq x_0$. We denote i_0 as the smallest i that satisfies (4). Since $x > x_0$, we get $x \geq 3\sum_{i=1}^{i_0}(t_i' + T_i')$. Then we divide the satisficing regret of `SELECT-LITE` into three parts: 1. the satisficing regret incurred in Step 1 and 2 by round i_0; 2. the satisficing regret incurred in Step 1 and 2 after round i_0; 3. the satisficing regret incurred in Step 3 of all rounds. Since $x \geq 3\sum_{i=1}^{i_0}(t_i' + T_i')$, the first part of satisficing regret is at most $x/3$. Therefore if the total satisficing regret exceeds x, then either the second or the third part of satisficing regret exceeds $x/3$.

To bound the tail distribution of the second part of satisficing regret, we first define k_0 as the maximum integer such that $\sum_{i=i_0+1}^{i_0+k_0}(t_i' + T_i') \leq x/3$. Then we verify that $k_0 = \Omega(x^{\zeta})$. If the second part of satisficing regret exceeds $x/3$, then the total number of rounds should be at least $i_0 + k_0 + 1$, and by Proposition 4, `SELECT-LITE` enters round $i_0 + k_0 + 1$ (*i.e.*, round $i_0 + k_0$ terminates within the time horizon) with probability at most $4^{-k_0} = \exp(-\Omega(x^{\zeta}))$. Therefore the probability that the second part of satisficing regret exceeds $x/3$ is at most $\exp(-\Omega(x^{\zeta}))$.

To bound the tail distribution of the third part of satisficing regret, by the definition of confidence intervals we have that the probability that the number of steps in the LCB test exceeds k is $O(\exp(-k^{\zeta}))$. Furthermore, by Proposition 4, the total number of rounds of `SELECT-LITE` is upper bounded by a shifted geometric random variable with success probability $3/4$. Then we use the following proposition to bound the tail distribution of the third part of satisficing regret.

Proposition 5. *Suppose $\{Q_i\}$ are independent nonnegative random variables and there exists $K_1 > 0$, $0 < \zeta < 1$ such that for all $x > 0$ and all $i \in \mathbb{Z}^+$,*

$$Pr(Q_i \geq x) \leq K_1 \exp(-x^\zeta).$$

Let $N(\lambda)$ be a geometric distribution with success rate being $1-\lambda$, where $\lambda < 1/2$. Then for all $x > 0$, we have

$$Pr\left(\sum_{i=1}^{N+N(\lambda)} Q_i \geq x\right) \leq \frac{2^{N+1}(1-\lambda)}{1-2\lambda} \exp\left(-\frac{x^\zeta}{K_1+1}\right).$$

Proposition 5 states that the sum of a random number of independent light-tailed random variables is still light-tailed if the number of random variables follows a shifted geometric distribution. Proof of Proposition 5 is provided in Section E.2. By Proposition 5, the probability that the total number of time steps for Step 3 exceeds $x/3$ is at most $\exp(-\Omega(x^\zeta))$. Therefore the probability that the satisficing regret incurred in Step 3 exceeds $x/3$ is at most $\exp(-\Omega(x^\zeta))$. Therefore we conclude that the probability that the satisficing regret of `SELECT-LITE` exceeds x is at most $\exp(-\Omega(x^\zeta))$ for large enough x. □

Remark 9 (**Independence of T for the Tail Probability Bound**). *We would like to point out that the tail probability bound of `SELECT-LITE` does not depend on T. In other words, for any realizable instance, there exists a fixed light-tailed distribution that dominates the distribution of satisficing regret of `SELECT-LITE` with any time horizon T. This is in sharp contrast with the tail probability bounds provided in prior works such as Simchi-Levi et al. (2025), which deteriorate as T increases although the standard regret distribution for every fixed T is light-tailed.*

We use the following theorem to show that `SELECT-LITE` enjoys a constant expected satisficing regret bound.

Theorem 7. *If $r(A^*) \geq S$, then the expected satisficing regret of `SELECT-LITE` is bounded by*

$$\mathbb{E}[\textit{Regret}_S] \leq \min\left\{C_1\left(\frac{C_1}{\Delta_S^*}\right)^{\frac{\zeta+\alpha(1-\zeta)}{(1-\zeta)^2(1-\alpha)}} \cdot \text{polylog}\left(\frac{C_1}{\Delta_S^*}\right), C_1 T^{\alpha(1-\zeta)+\zeta} \cdot \text{polylog}(T)\right\}.$$

Proof of Theorem 7 is similar to that of Theorem 1 and is provided in Section F of the appendix. Theorem 7 states that for all $0 < \zeta < 1$, `SELECT-LITE` attains a constant expected satisficing regret bound depending on ζ. The power of $1/\Delta_S^*$ in the expected satisficing regret bound is monotonically increasing in ζ, suggesting that stronger guarantees on tail distribution of satisficing regret leads to larger constant satisficing regret on its expectation. When ζ approaches zero, the power of $1/\Delta_S^*$ approaches $\alpha/(1-\alpha)$, which is the same as the one in the expected satisficing regret bound of `SELECT`. This implies that when ζ is small, `SELECT-LITE` can have a similar expected satisficing regret as `SELECT`.

We would like to remark that although both SELECT and SELECT-LITE are able to attain a constant expected satisficing regret bound, the expected satisficing regret bound of SELECT-LITE is always weaker than that of SELECT. Specifically, the constant bound of SELECT-LITE scales with C_1/Δ_S^* to the power of $(\zeta+\alpha(1-\zeta))/(1-\zeta)^2(1-\alpha)$, while the constant bound of SELECT scales with C_1/Δ_S^* to the power of $\alpha/(1-\alpha)$. Since $0<\zeta<1$ and $0<\alpha<1$, we have $\zeta+\alpha(1-\zeta)>\alpha$ and $(1-\zeta)^2(1-\alpha)<1-\alpha$. Therefore we always have $(\zeta+\alpha(1-\zeta))/(1-\zeta)^2(1-\alpha)>\alpha/(1-\alpha)$, implying that the constant bound of SELECT-LITE is always weaker than that of SELECT. Numerical experiments also show that SELECT outperforms SELECT-LITE in terms of expected satisficing regret, which we refer to Section 7.4 for details. To conclude, although SELECT-LITE provides stronger protection against tail risk, such protection comes at the cost of a higher expected satisficing regret compared to SELECT. In applications that are sensitive towards tail risk, it would be better for the decision maker to use SELECT-LITE. On the other hand, if the decision maker only cares about satisficing regret in expectation, than SELECT is able to better serve the purpose of minimizing expected satisficing regret.

Finally, we show that in the non-realizable case, the expected standard regret of SELECT-LITE is sub-linear in T.

Theorem 8. *If $r(A^*)<S$, the expected standard regret of SELECT-LITE is bounded by*

$$\mathbb{E}[\textit{Regret}] \leq C_1 T^{\alpha(1-\zeta)+\zeta} \cdot \text{polylog}(T).$$

Proof of Theorem 8 is provided in Section G of the appendix. Theorem 8 states that for all $0<\zeta<1$, SELECT-LITE always attains a sub-linear in T expected standard regret bound in the non-realizable case. When ζ approaches zero, the power of T in the expected standard regret bound of SELECT-LITE approaches α, which is the power of T in the expected standard regret bound of SELECT and the learning oracle used in Step 1. In other words, when ζ is small, SELECT-LITE can have a similar expected standard regret as SELECT.

6.4. Novelty of the Results

In this section, we would like to highlight some major novelties of our algorithm SELECT-LITE.

New Approach towards Light-Tailed Regret Distribution: In SELECT-LITE, a light-tailed satisficing regret distribution is attained mainly from a novel multi-round structure of the algorithm, where we gradually increase the number of time steps of running the learning oracle to search for satisficing arms. In Proposition 4, we prove that the number of rounds of SELECT-LITE is upper bounded by a shifted geometric distribution, which implies that the probability of starting a new round decays exponentially as the rounds proceed. With a polynomial growth in the length of running the oracle and an exponential decay in the probability of starting a new round, we are able

to prove that the total number of time steps of running the oracle and forced sampling. This design is novel in the literature of tail risk control in bandits. Our approach is also fundamentally different from existing ones in Fan and Glynn (2024) and Simchi-Levi et al. (2025), where a light-tailed regret distribution is obtained by using larger confidence intervals that hold with a probability of at least $1-\exp(-\Omega(T^{\varsigma}))$.

Nuanced Difference between Satisficing Regret and Standard Regret: Simchi-Levi et al. (2025) show that for standard regret minimization in finite-armed bandits, any algorithm with a light-tailed standard regret distribution will inevitably incur a polynomial in T instance-dependent expected standard regret. In other words, no algorithm can maintain both a light-tailed standard regret distribution and a near-optimal $O(\log(T))$ instance-dependent expected standard regret bound at the same time. Because the constant expected satisficing regret bounds are typically instance-dependent, it becomes unclear if a constant expected satisficing regret is compatible with a light-tailed satisficing regret distribution. In this paper, we show that it is indeed possible to break this impossible regime. Specifically, `SELECT-LITE` is able to attain a light-tailed satisficing regret distribution and a constant instance-dependent expected satisficing regret bound in the realizable case. Furthermore, our results can be extended beyond finite-armed bandits and are applicable to any class of bandit optimization problems that has a learning oracle with a sub-linear expected standard regret bound. This reveals a fundamental difference between the notions of standard regret and satisficing regret.

No Additional Assumptions on the Learning Oracle: We emphasize that our results do not rely on any condition on the learning oracle other than a sub-linear expected standard regret (Condition 1). In particular, we do not require that the standard regret distribution of the learning oracle is light-tailed. Instead, `SELECT-LITE` is able to make use of learning oracles with a heavy-tailed standard regret distribution to attain a light-tailed satisficing regret distribution in the realizable case. For example, the satisficing regret distribution of `SELECT-LITE` is light-tailed when using the UCB algorithm or Thompson sampling as the learning oracle in Step 1, even though both algorithms are proven to have a heavy-tailed standard regret distribution (see Fan and Glynn 2024). The reason why the tail distribution of the learning oracle's standard regret does not change the result is that with the help of Proposition 4, we are able to directly prove that the total number of time steps of running the learning oracle and forced sampling follows a light-tailed distribution. Since we have assumed that the mean reward of each arm belongs to $[0,1]$, the total satisficing regret incurred when running the learning oracle and forced sampling follows a light-tailed distribution.

The ability of making use of learning oracles with potentially heavy-tailed standard regret distributions is important for `SELECT-LITE` to generalize to a broader range of bandit optimization problem

classes. To the best of our knowledge, algorithms with light-tailed standard regret distribution have been established for only a very limited range of bandit optimization problems including finite-armed bandits and linear bandits. By accepting algorithms with potentially heavy-tailed standard regret distribution as learning oracle, `SELECT-LITE` is able to extend beyond this scope to any class of bandit optimization problems that admits a learning oracle with sub-linear expected standard regret bound.

Remark 10 (**Light-Tailed Standard Regret in the Non-Realizable Case**). *We would like to clarify that although we have proven a light-tailed satisficing regret distribution for `SELECT-LITE` in the realizable case, `SELECT-LITE` may not be able to attain a light-tailed standard regret distribution in the non-realizable case. The main reason for this is that we make no assumption on the standard regret distribution of the learning oracle. Because in the non-realizable case, `SELECT-LITE` could run the learning oracle in $\Theta(T)$ time steps, it could have a heavy-tailed standard regret distribution in the non-realizable case if the standard regret distribution of the learning oracle is heavy-tailed. In Section H of the appendix, we introduce an algorithm for finite-armed bandits that, besides achieving a constant expected satisficing regret, attains both a light-tailed satisficing regret distribution in the realizable case and a light-tailed standard regret distribution in the non-realizable case.*

7. Numerical Experiments on Synthetic Data

In this section, we conduct numerical experiments to test the performance of `SELECT` on finite-armed bandits, concave bandits, and Lipschitz bandits.

7.1. Finite-Armed Bandits

Setup: In this case, we consider an instance of 4 arms, and the expected rewards of all arms are $\{0.2, 0.4, 0.6, 0.8\}$. We vary the length of the time horizon from 500 to 7000 with a stepsize of 500. We conduct experiments for both the realizable case and the non-realizable case. For the realizable case, we set the satisficing level $S = 0.7$; for the non-realizable case, we set the satisficing level $S = 1.5$. In both cases, we compare `SELECT` against the "Phase Exploration with Bounded Regret" algorithm[5] in Garivier et al. (2019) (referred as Phase Exploration in the following), STS algorithm in Russo and Van Roy (2022) with adaptations to the time-undiscounted setting, Thompson sampling in Agrawal and Goyal (2017) as well as SAT-UCB and SAT-UCB+ in Michel et al. (2023). Note that in the setting of Russo and Van Roy (2022), a satisficing action (arm) A is defined as a near-optimal one, *i.e.*, $r(A) \geq r(A^*) - D$ with given $D > 0$, and the regret of a single pull is defined as $r(A^*) - r(A) - D$ which can be negative, which is different to our definition on satisficing regret. Here, we run STS directly with our satisficing level $S = r(A^*) - D$. For our

[5] Algorithm 1, Appendix C in Garivier et al. (2019), with μ^* changed to S.

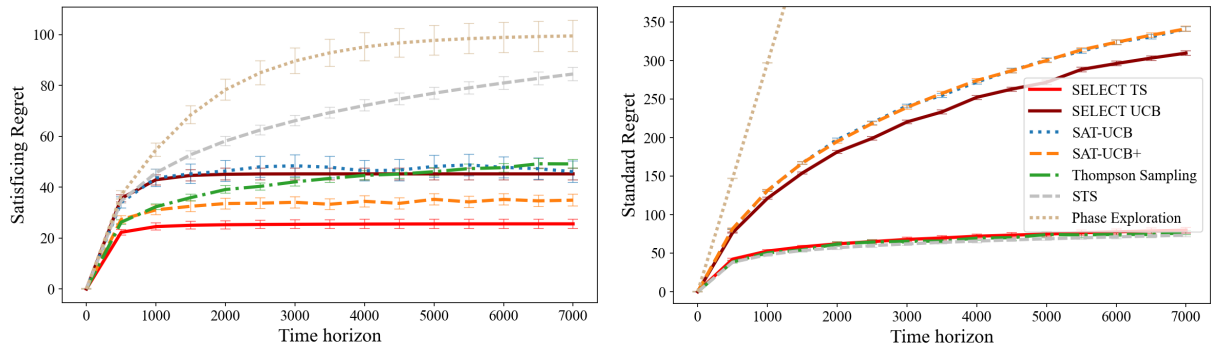
algorithm SELECT, we use SELECT-BoBW, a variant of SELECT that attains a best-of-both-worlds constant satisficing regret bound for finite-arm bandits (we refer to Algorithm 4 in Section N of the appendix for details), and use UCB algorithm (Bubeck and Cesa-Bianchi 2012) and Thompson sampling respectively as the learning oracle used in Step 1. The experiment is repeated for 5000 times and we report the average satisficing regret in the realizable case and the average standard regret in the non-realizable case.

Results: The results of the realizable case are presented in Figure 2(a), and the results of the non-realizable case are presented in Figure 2(b). From Figure 2(a), one can see that the expected satisficing regret of SELECT with both oracle algorithms, SAT-UCB, and SAT-UCB+ barely increases when T exceeds 2500, and the expected satisficing regret of Phase Exploration algorithm barely increases when T exceeds 6000. This suggests that SELECT, SAT-UCB and SAT-UCB+ are able to attain a constant expected satisficing regret in the realizable case, consistent with the corresponding theoretical results. On the other hand, both STS and Thompson sampling fail to exhibit a constant expected satisficing regret in the realizable case. When comparing the expected satisficing regret of these algorithms, one can see that SELECT using Thompson sampling as oracle algorithm incurs the smallest satisficing regret among all tested algorithms, and the expected satisficing regret of SELECT using UCB as the oracle algorithm is comparable with that of SAT-UCB, slightly outperformed by the heuristic SAT-UCB+, and is still lower than Thompson sampling, STS, and the Phase Exploration algorithm. Although Phase Exploration algorithm eventually converges to a constant expected satisficing regret, it converges slowly and incurs a high satisficing regret among all tested algorithms.

From Figure 2(b) one can see that the expected standard regret of SELECT using Thompson sampling as oracle algorithm in the non-realizable case is comparable to that of Thompson sampling and STS. The expected standard regret of SELECT using UCB as oracle algorithm in the non-realizable case is higher, but SELECT with both oracle algorithms incurs smaller standard regret than SAT-UCB and SAT-UCB+. One can also see that Phased Exploration fails in the non-realizable setting, whose expected standard regret scales linearly as T increases.

7.2. Concave Bandits

Setup: In this case, we consider an instance with arm set $[0,1]$ and concave reward function $r(x) = 1 - 16(x - 0.25)^2$. We vary the length of the time horizon from 500 to 5000 with a stepsize of 500. We consider both the realizable case and the non-realizable case. For the realizable case, we set the satisficing level $S = 0.3$; for the non-realizable case, we set the satisficing level $S = 1.5$. For both cases, we compare our algorithm SELECT with the convex bandit algorithm introduced in Agarwal et al. (2011). For SELECT, we use the algorithm in Agarwal et al. (2011) as the learning oracle.



(a) Satisficing Regret in the Realizable Case (b) Standard Regret in the Non-Realizable Case

Figure 2 Comparison of `SELECT-BoBW`, Thompson sampling, STS, Phase Exploration, SAT-UCB and SAT-UCB+

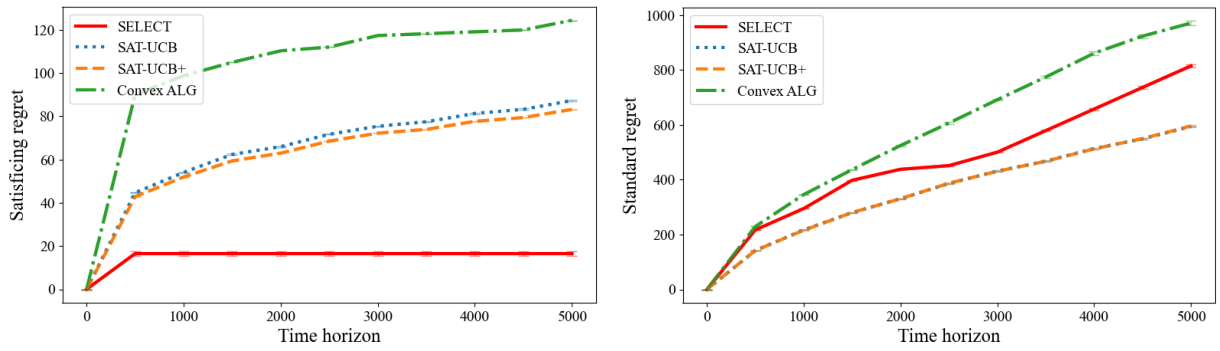
We also use SAT-UCB and SAT-UCB+ as heuristics by viewing the problem as Lipschitz bandits. That is, we first discretize the arm set uniformly with stepsize $L^{-2/3}T^{-1/3}$, where L is the Lipschitz constant of the reward function, and then run SAT-UCB and SAT-UCB+ over the discretized arm set. Note that for SAT-UCB and SAT-UCB+, we use a discretization stepsize that decreases as T increases instead of a constant stepsize. The reason why we do not use a constant stepsize is that for any fixed constant discretization stepsize, there always exists an realizable instance such that none of the discretized arms are satisficing. This observation suggests that SAT-UCB and SAT-UCB+ with any constant discretization stepsize could incur linear in T satisficing regret even when satisficing arms exist. The experiment is repeated for 1000 times, and we report the average results.

Results: The results of realizable case are provided in Figure 3(a), and the results of non-realizable case are provided in Figure 3(b). One can see from Figure 3(a) that the expected satisficing regret of `SELECT` becomes stable after T exceeds 500, suggesting that `SELECT` attains a constant expected satisficing regret in the realizable case. On the other hand, all the other three algorithms fail to exhibit a constant expected satisficing regret in the realizable case. When comparing the expected satisficing regret of the four algorithms, we observe that `SELECT` incurs a smaller expected satisficing regret than all three other algorithms.

One can see from Figure 3(b) that in the non-realizable case, `SELECT` incurs a smaller expected standard regret compared to the algorithm in Agarwal et al. (2011), while performance of `SELECT` is only slightly worse than both SAT-UCB and SAT-UCB+.

In this experiment, the reason why `SELECT` outperforms the algorithm in Agarwal et al. (2011) can be explained as follows. In the instance we use in this experiment, both the set of optimal arms and the set of significantly suboptimal arms are large, and directly running the algorithm in Agarwal et al. (2011) over the entire time horizon tend to over-explore due to larger confidence

intervals, thus would potentially incur significant regret. In contrast, SELECT starts with running the algorithm in Agarwal et al. (2011) over a small number of time steps t_i. This allows SELECT to have a much smaller confidence intervals in early stages of the algorithm, thus explores less and exploits more aggressively. Furthermore, when running forced sampling in Step 2, SELECT also exploits a reasonably good arm obtained from running the learning oracle, which means even more exploitation for SELECT. Therefore, although theoretically SELECT may underperform Agarwal et al. (2011) by a logarithmic factor of T, in the instance used in Figure 3(b), SELECT outperforms the algorithm in Agarwal et al. (2011). Similar phenomenon can also be observed in the numerical experiment for Lipschitz bandits in Section 7.3 and for linear bandits in Section 8.2.



(a) Satisficing Regret in the Realizable Case

(b) Standard Regret in the Non-Realizable Case

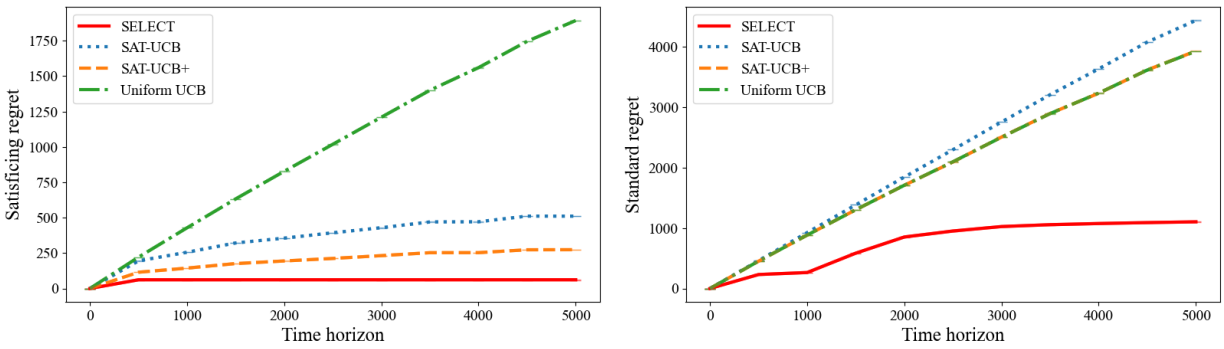
Figure 3 Comparison of SELECT, Stochastic Convex, SAT-UCB and SAT-UCB+ for concave bandit

7.3. Lipschitz Bandits

Setup: In this case, we consider a two-dimensional Lipschitz bandit in domain $(x, y) \in [0, 1]^2$ with Lipschitz rewards $r(x, y) = \min\{1, 3e^{-100((x-0.5)^2+(y-0.7)^2)}\}$. We vary the length of the time horizon from 500 to 5000 with a stepsize of 500. We consider both the realizable case and the non-realizable case. For the realizable case, we set the satisficing level S = 0.5; for the non-realizable case, we set the satisficing level S = 1.5. For both cases, we compare our algorithm SELECT with the uniformly discretized UCB introduced in Bubeck et al. (2011) (hereafter referred to as the "Uniform UCB"). For SELECT, we use the Uniform UCB as the learning oracle. We again use SAT-UCB and SAT-UCB+ as heuristics by uniformly discretizing the arm set with stepsize $L^{-1/2}T^{-1/4}$, where L is the Lipschitz constant of the reward function, and run the algorithms over the discretized arms. Note that for SAT-UCB and SAT-UCB+, we use a discretization stepsize that decreases as T increases instead of a constant stepsize. The reason for this is that, as we have mentioned in Section 7.2, SAT-UCB and SAT-UCB+ with constant stepsize could incur a satisficing regret that is linear in T. The experiment is repeated for 1000 times and we report the average results.

Results: The results of realizable case are provided in Figure 4(a), and the results of non-realizable case are provided in Figure 4(b). One can see from Figure 4(a) that the expected satisficing regret of `SELECT` becomes stable after T exceeds 500, suggesting that `SELECT` attains a constant expected satisficing regret in the realizable case. On the other hand, none of the other three algorithms are able to attain a constant expected satisficing regret. When comparing the expected satisficing regret of the four algorithms, one can observe that `SELECT` incurs an expected satisficing regret smaller than all three other algorithms.

As for the non-realizable case, one can see from Figure 4(b) that `SELECT` achieves the best empirical performance among the four algorithms, while the empirical performance of SAT-UCB and SAT-UCB+ is similar to that of Uniform UCB in the non-realizable case.



(a) Satisficing Regret in the Realizable Case (b) Standard Regret in the Non-Realizable Case

Figure 4 Comparison of `SELECT`, Uniform UCB, SAT-UCB and SAT-UCB+ for Lipschitz bandit

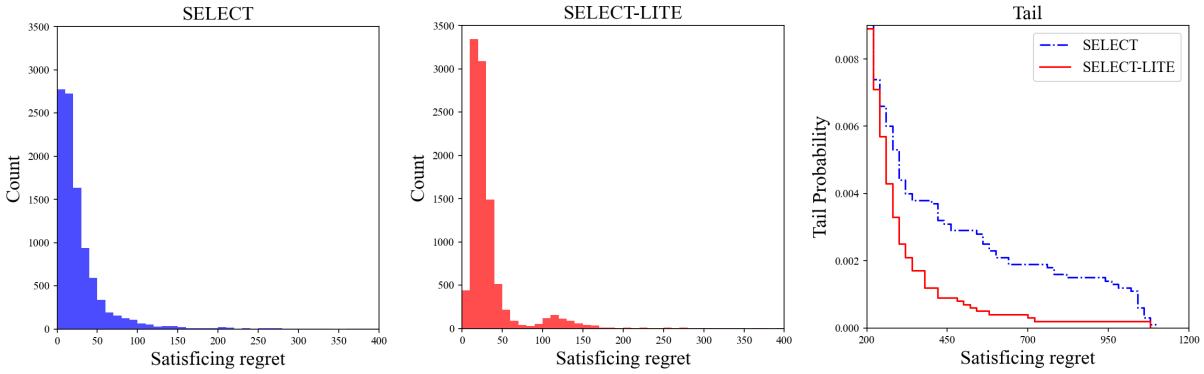
7.4. Tail Risk Control in Finite-Armed Bandits

We test the performance of `SELECT-LITE` in terms of both expected satisficing regret and tail distribution of satisficing regret using an instance of finite-armed bandit.

Setup: We consider an instance of 4 arms. The expected rewards of all arms are set as $\{0.2, 0.4, 0.6, 0.8\}$. For the realizable case, we set the satisficing level $S = 0.7$; for the non-realizable case, we set the satisficing level $S = 1.5$. For `SELECT` and `SELECT-LITE`, we use Thompson sampling as the learning oracle. We vary $\zeta \in \{0.1, 0.4\}$.

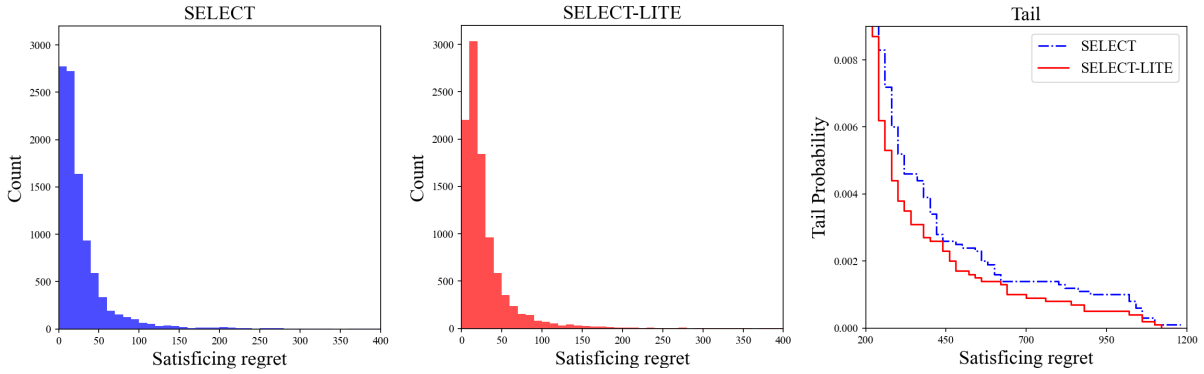
We first compare the expected satisficing regret of `SELECT-LITE` with that of `SELECT` and Thompson sampling. In this case, we vary the length of the time horizon from 1000 to 10000 with a stepsize of 1000. The experiment is repeated for 1000 times and we report the average satisficing regret. Then we compare the empirical tail distribution of satisficing regret between `SELECT-LITE` and `SELECT` under time horizon $T = 10000$. The experiment is repeated for 10000 times.

Results: The results of realized satisficing regret distribution under $\zeta = 0.1$ are presented in Figure 5(a) for `SELECT` and Figure 5(b) for `SELECT-LITE` , and the comparison on tail probability is presented in Figure 5(c) by zooming in on the tail region of the distributions. The results under $\zeta = 0.4$ are presented in Figure 6. Here we report the empirical distribution of realized satisficing regret using histograms with a bin width of 10. From Figure 5(c) and 6(c), one can see that under both $\zeta = 0.1$ and $\zeta = 0.4$, `SELECT-LITE` exhibits a lower tail probability compared to `SELECT`. When comparing between between the case of $\zeta = 0.1$ and the case of $\zeta = 0.4$, `SELECT-LITE` with $\zeta = 0.1$ exhibits a noticeably thinner tail, even though a larger ζ should theoretically yield the faster decay rate. This is because when $\zeta = 0.1$, t_i' increases rapidly in the first few rounds, thus `SELECT-LITE` can quickly find the satisficing arm within a few rounds. Consequently, the empirical tail of `SELECT-LITE` when $\zeta = 0.4$ is a summation of tails of many rounds, but the one when $\zeta = 0.1$ is a summation of simply one or two rounds, which potentially makes the tail probability smaller.



(a) `SELECT` Regret Distribution (b) `SELECT-LITE` Regret Distribution (c) Tail Comparison

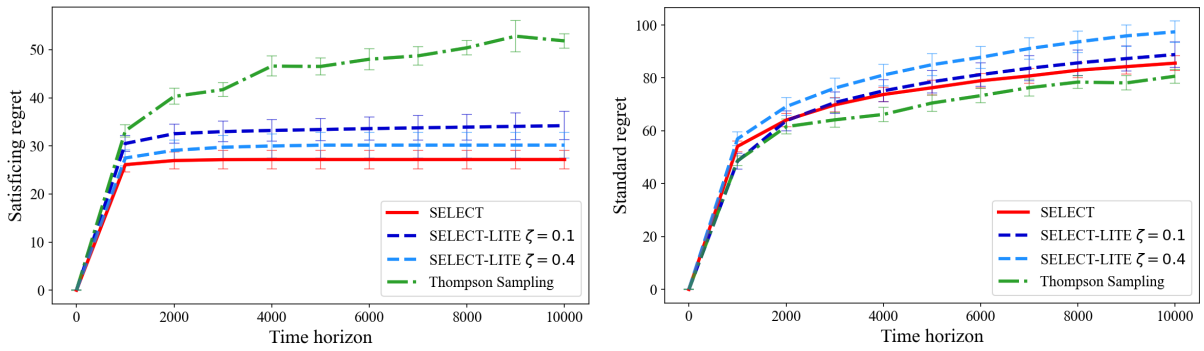
Figure 5 Realized satisficing regret distribution of `SELECT` and `SELECT-LITE` at $T = 10000$ with $\zeta = 0.1$



(a) `SELECT` Regret Distribution (b) `SELECT-LITE` Regret Distribution (c) Tail Comparison

Figure 6 Realized satisficing regret distribution of `SELECT` and `SELECT-LITE` at $T = 10000$ with $\zeta = 0.4$

The results of expected satisficing regret in the realizable case are presented in Figure 7(a). The results of expected standard regret in the non-realizable case are presented in Figure 7(b). From Figure 7(a) one can observe that `SELECT-LITE` exhibits constant satisficing regret that is only slightly larger than that of `SELECT`. Furthermore, both when $\zeta = 0.1$ and when $\zeta = 0.4$, `SELECT-LITE` incur less satisficing regret compared to Thompson sampling. From Figure 7(b) one can see that `SELECT-LITE` with different ζ incurs roughly the same standard regret as `SELECT` and is slightly worse than Thompson sampling. This is because each `SELECT-LITE` algorithm shares a backbone with `SELECT`, and `SELECT-LITE` with $\zeta = 0.1$ has a $O(T^{0.55})$ regret bound that is very close to that of `SELECT`. For `SELECT-LITE` with $\zeta = 0.4$, although the theoretical standard regret is $O(T^{0.7})$, worse than the $O(T^{0.5})$ guarantee for Thompson sampling. With these observations, we conclude that by running `SELECT-LITE` instead of `SELECT`, the decision maker can have better protection against tail risks, both theoretically and empirically, without incurring a significantly larger satisficing regret in expectation.



(a) Satisficing Regret in the Realizable Case (b) Standard Regret in the Non-Realizable Case

Figure 7 Comparison of `SELECT` , `SELECT-LITE` with different ζ, and Thompson sampling

8. Numerical Experiments on Real Data: Satisficing in Dynamic Pricing

To further evaluate the empirical performance of our algorithm in more practical settings, in this section, we conduct numerical experiments on an instance calibrated from the avocado dataset Kiggins (2018).

8.1. Data Description, Pre-Processing, and Setup

We use the avocado dataset Kiggins (2018) in the New York region. The data records weekly average prices of organic avocado and the corresponding weekly total selling volume from January 4, 2015, to December 27, 2015, for a total of 52 weeks (with outliers deleted). The weekly average prices in these 52 data points range from \$1.65 to \$2.34, and the average price is \$2.03. The weekly selling volumes range from 9205 to 25675, and the average weekly selling volume is 17106.

We consider a dynamic pricing problem with linear demands to the prices, and test the performance of the algorithms on this model fitted by the avocado dataset. Assuming that the demand is equal to $D_t = g - hp_t + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2)$, given price p_t at each time step t, we fit the linear regression model from the real data to be $\hat{g} = 32724$, $\hat{h} = 7678$, $\hat{\sigma} = 4100$. By normalizing the model to ensure unit noise variance, we obtain the coefficients to be

$$\hat{g}_n = 32724/4100,\ \hat{h}_n = 7678/4100,\ \hat{\sigma}_n = 1.$$

We also extend the range of average price to be $p \in [0, 4]$. The mean revenue received at each time t is $p_t(\hat{g}_n - \hat{h}_n p_t)$, and the mean revenue of each feasible price is visualized in Figure 8.

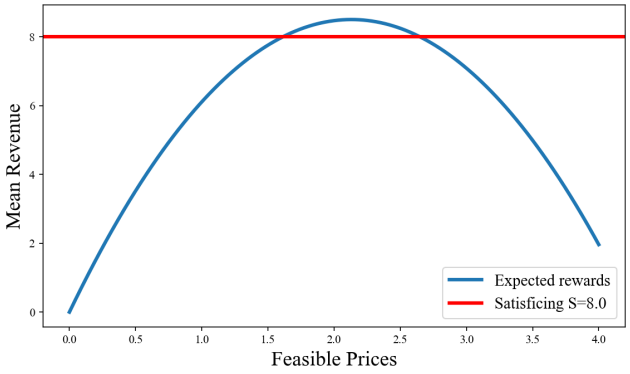
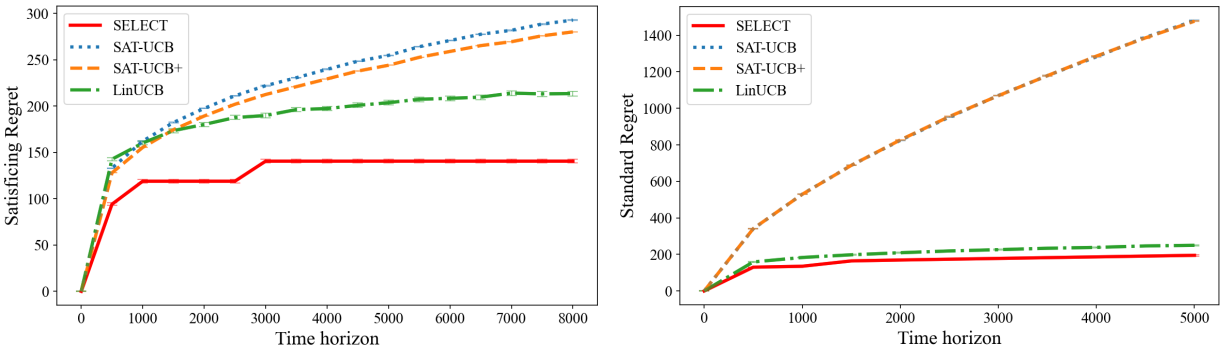


Figure 8 **Expected revenue of the feasible prices** $p \in [0, 4]$

We set the length of the time horizon to be 500 to 8000 for the realizable case and 5000 for the non-realizable case with a stepsize of 500. For the realizable case, we set the satisficing level $S = 8.0$; for the non-realizable case, we set the satisficing level $S = 10.0$. For both cases, we compare our algorithm `SELECT` with Linear UCB introduced in Lattimore and Szepesvári (2020). For `SELECT`, we use Linear UCB in Lattimore and Szepesvári (2020) as the learning oracle. We also use SAT-UCB and SAT-UCB+ as heuristics by viewing the problem as Lipschitz bandits. That is, we first discretize the arm set uniformly with a stepsize $L^{-1/3}T^{-1/3}$, where L is the Lipschitz constant of the reward function, and then run SAT-UCB and SAT-UCB+ over the discretized arm set. The experiment is repeated 1000 times and we report the averaged results.

8.2. Results

The results of the realizable case are provided in Figure 9(a), and the results of the non-realizable case are provided in Figure 9(b). From Figure 9(a), one can observe that the expected satisficing regret of `SELECT` becomes stable after T exceeds 3000. This suggests that in the realizable case, `SELECT` is able to attain constant expected satisficing regret. On the other hand, none of the other three algorithms are able to achieve constant expected satisficing regret in the realizable case. When



(a) Satisficing Regret in the Realizable Case (b) Standard Regret in the Non-Realizable Case

Figure 9 Comparison of `SELECT` , Linear UCB, SAT-UCB and SAT-UCB+ for linear bandit

comparing the expected satisficing regret of the four algorithms, one can see that `SELECT` incurs a smaller expected satisficing regret than all the other three algorithms.

From Figure 9(b), one can see that in the non-realizable case, `SELECT` slightly outperforms Linear UCB and significantly outperforms both SAT-UCB and SAT-UCB+. The reason why `SELECT` and Linear UCB significantly outperforms SAT-UCB and SAT-UCB+ is that in the problem, SAT-UCB and SAT-UCB+ views the problem as a Lipschitz bandit, without making use of the linear structure of demand. The reason why `SELECT` outperforms Linear UCB in terms of standard regret is the same as that observed in numerical experiments on concave bandits in Section 7.2

9. Conclusion

In this paper, we propose `SELECT`, a general algorithmic template, for satisficing regret minimization in bandit optimization. For a given class of bandit optimization problems and a sub-linear regret learning oracle, we show that `SELECT` attains constant expected satisficing regret in the realizable case and the same regret as the learning oracle in the non-realizable case. We instantiate `SELECT` to finite-armed bandits, concave bandits, and Lipschitz bandits, and also make a discussion on the corresponding lower bounds. We also consider tail risk control in satisficing regret minimization, where we introduce `SELECT-LITE`, which attains a light-tailed satisficing regret distribution while still maintaining a constant expected satisficing regret. Finally, we conduct numerical experiments to validate the performance of `SELECT` and `SELECT-LITE` on both synthetic datasets and a real-world dynamic pricing case study.

For future studies, we suggest further exploration in the following directions. First, our work mainly focuses on the stationary settings, and further research could explore satisficing in the nonstationary online learning settings (Wei and Luo 2021, Cheung et al. 2022). Furthermore, the notion of satisficing regret indeed discourages exploration. Therefore, another direction is to study the role of satisficing regret in settings with limited adaptivity (Gao et al. 2019, Feng et al. 2025).

References

Abernethy, Jacob, Kareem Amin, Ruihao Zhu. 2016. Threshold bandits, with and without censored feedback. *Advances in Neural Information Processing Systems* **29**.

Agarwal, Alekh, Dean P Foster, Daniel J Hsu, Sham M Kakade, Alexander Rakhlin. 2011. Stochastic convex optimization with bandit feedback. *Advances in Neural Information Processing Systems* **24**.

Agrawal, Shipra, Navin Goyal. 2017. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)* **64**(5) 1–24.

Anonymous. 2025. Satisficing regret minimization in bandits. *The Thirteenth International Conference on Learning Representations*.

Audibert, Jean-Yves, Sébastien Bubeck. 2010. Best arm identification in multi-armed bandits. *COLT-23th Conference on learning theory-2010*. 13–p.

Audibert, Jean-Yves, Rémi Munos, Csaba Szepesvári. 2009. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science* **410**(19) 1876–1902.

Bubeck, Sébastien, Nicolo Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* **5**(1) 1–122.

Bubeck, Sébastien, Vianney Perchet, Philippe Rigollet. 2013. Bounded regret in stochastic multi-armed bandits. *Conference on Learning Theory*. PMLR, 122–134.

Bubeck, Sébastien, Gilles Stoltz, Jia Yuan Yu. 2011. Lipschitz bandits without the lipschitz constant. *International Conference on Algorithmic Learning Theory*. Springer, 144–158.

Caplin, Andrew, Mark Dean, Daniel Martin. 2011. Search and satisficing. *American Economic Review* **101**(7) 2899–2922.

Cheung, Wang Chi, David Simchi-Levi, Ruihao Zhu. 2022. Hedging the drift: Learning to optimize under nonstationarity. *Management Science* **68**(3) 1696–1713.

Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. 2009. *Introduction to Algorithms*. MIT Press.

Fan, Lin, Peter W Glynn. 2024. The fragility of optimized bandit algorithms. *Operations Research* .

Feng, Qing, Ruihao Zhu, Stefanus Jasin. 2025. Temporal fairness in learning and earning: Price protection guarantee and phase transitions. *Operations Research* **73**(2) 775–797.

Fitzsimons, Gavan J. 2000. Consumer response to stockouts. *Journal of Consumer Research* **27**(2) 249–266.

Gao, Zijun, Yanjun Han, Zhimei Ren, Zhengqing Zhou. 2019. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems* **32**.

Garivier, Aurélien, Pierre Ménard, Gilles Stoltz. 2019. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research* **44**(2) 377–399.

Google. 2026. Concepts in service monitoring URL `https://docs.cloud.google.com/stackdriver/docs/solutions/slo-monitoring`.

Hajiabolhassan, Hossein, Ronald Ortner. 2025. Online regret bounds for satisficing in markov decision processes. INFORMS.

Hüyük, Alihan, Cem Tekin. 2021. Multi-objective multi-armed bandit with lexicographically ordered and satisficing objectives. *Machine Learning* **110**(6) 1233–1266.

Jing, Xiaoqing, Michael Lewis. 2011. Stockouts in online retailing. *Journal of Marketing Research* **48**(2) 342–354.

Kano, Hideaki, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, Masashi Sugiyama. 2019. Good arm identification via bandit feedback. *Machine Learning* **108** 721–745.

Kiggins, Justin. 2018. Avocado prices: Historical data on avocado prices and sales volume in multiple us markets. `https://www.kaggle.com/datasets/neuromusic/avocado-prices`.

Lattimore, Tor, Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

Levy, Michael, Barton Weitz, Dhruv Grewal. 2023. *Retailing Management*. McGraw Hill.

Locatelli, Andrea, Maurilio Gutzeit, Alexandra Carpentier. 2016. An optimal algorithm for the thresholding bandit problem. *International Conference on Machine Learning*. PMLR, 1690–1698.

Lufkin, Bryan. 2021. Do 'maximisers' or 'satisficers' make better decisions? *BBC* .

Mabey, Chris. 2022. Optimizing vs satisficing: Tips for product design and designing your life. *The BYU Design Review* .

Michel, Thomas, Hossein Hajiabolhassan, Ronald Ortner. 2023. Regret bounds for satisficing in multi-armed bandit problems. *Transactions on Machine Learning Research* .

Mukherjee, Subhojyoti, Naveen Kolar Purushothama, Nandan Sudarsanam, Balaraman Ravindran. 2017. Thresholding bandits with augmented ucb. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*. 2515–2521.

Reverdy, Paul, Naomi E Leonard. 2014. Satisficing in gaussian bandit problems. *53rd IEEE Conference on Decision and Control*. IEEE, 5718–5723.

Reverdy, Paul, Vaibhav Srivastava, Naomi Ehrich Leonard. 2017. Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control* **62**(8) 3788–3803.

Russo, Daniel, Benjamin Van Roy. 2022. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research* **47**(4) 2815–2839.

Salomon, Antoine, Jean-Yves Audibert. 2011. Deviations of stochastic bandit regret. *International Conference on Algorithmic Learning Theory*. Springer, 159–173.

Simchi-Levi, David, Zeyu Zheng, Feng Zhu. 2023. Stochastic multi-armed bandits: optimal trade-off among optimality, consistency, and tail risk. *Advances in Neural Information Processing Systems* **36** 35619–35630.

Simchi-Levi, David, Zeyu Zheng, Feng Zhu. 2025. A simple and optimal policy design with safety against heavy-tailed risk for stochastic bandits. *Management Science* **71**(7) 6298–6318.

Simon, Herbert A. 1956. Rational choice and the structure of the environment. *Psychological Review* .

Slivkins, Aleksandrs, et al. 2019. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* **12**(1-2) 1–286.

Sloss, Benjamin Treynor. 2016. *Site Reliability Engineering*. O'Reilly.

Tamatsukuri, Akihiro, Tatsuji Takahashi. 2019. Guaranteed satisficing and finite regret: Analysis of a cognitive satisficing value function. *Biosystems* **180** 46–53.

Wei, Chen-Yu, Haipeng Luo. 2021. Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach. *Conference on Learning Theory* **134** 4300–4354.

Whitenton, Kathryn. 2014. Satisficing: Quickly meet users' main needs. *Nielsen Norman Group* .

Appendix

A. Proof of Theorem 1

Denote $\beta' = \max\{\beta, 1/2\}$ and $K_{\alpha,\beta} = \dfrac{32}{(1-\alpha)^\beta}$, and i_0 as

$$i_0 = \max\{i \in \mathbb{Z}_+ : C_1 K_{\alpha,\beta} \gamma_i \log(4/\gamma_i)^{\max\{\beta,1/2\}} > \Delta_S^*\}.$$

We first prove an upper bound on i_0. Denote $\delta_0 \in (0,1)$ the smallest solution to the following equation

$$\delta_0 \log(4/\delta_0)^{\beta'} = \frac{\Delta_S^*}{C_1 K_{\alpha,\beta}}. \tag{5}$$

Note that $\delta_0 \log(4/\delta_0)^{\beta'}$ is continuous on $(0,1)$. When $\delta_0 \to 0$, $\delta_0 \log(4/\delta_0)^{\beta'} \to 0$, and its maximum is $\max_{\delta_0} \delta_0 \log(4/\delta_0)^{\beta'} \geq \delta_0 \log(4/\delta_0)^{\beta'}\big|_{\delta_0 = 1^-} > 1$. On the other hand, since by definition $K_{\alpha,\beta} = 32/(1-\alpha)^\beta \geq 32$, the right hand side is upper bounded by $\Delta_S^*(1-\alpha)^\beta/(32C_1) \leq \Delta_S^*/32 \leq 1/32$ with $\Delta_S^* \leq 1$ and $1/2 \leq \alpha < 1$. Therefore there always exists $\delta_0 \in (0,1)$ such that (5) holds. Based on (5), we also have

$$\begin{aligned}
\frac{1}{\delta_0} =& \frac{C_1 K_{\alpha,\beta} \log(4/\delta_0)^{\beta'}}{\Delta_S^*} \\
=& \frac{C_1 K_{\alpha,\beta} \cdot (4\beta')^{\beta'} \log((4/\delta_0)^{1/(4\beta')})^{\beta'}}{\Delta_S^*} \\
\leq& \frac{C_1 K_{\alpha,\beta} \cdot (4\beta')^{\beta'} (4/\delta_0)^{1/4}}{\Delta_S^*} \\
=& \frac{C_1 K_{\alpha,\beta} \cdot (4\beta')^{\beta'} 2^{1/2}}{\Delta_S^* (\delta_0)^{1/4}} \\
\leq& \frac{(8\beta')^{\beta'} C_1 K_{\alpha,\beta}}{\Delta_S^* (\delta_0)^{1/4}}.
\end{aligned}$$

Here, the first inequality is because $\log(x) < x$ for all $x > 0$, and the second inequality is because $\beta' = \max\{\beta, 1/2\} \geq 1/2$, thus $2^{1/2} \leq 2^{\beta'}$. Therefore we get

$$\frac{1}{\delta_0} \leq \left(\frac{(8\beta')^{\beta'} C_1 K_{\alpha,\beta}}{\Delta_S^*} \right)^{4/3},$$

or equivalently,

$$\frac{4}{\delta_0} \leq 4 \left(\frac{(8\beta')^{\beta'} C_1 K_{\alpha,\beta}}{\Delta_S^*} \right)^{4/3} \leq \left(\frac{4(8\beta')^{\beta'} C_1 K_{\alpha,\beta}}{\Delta_S^*} \right)^{4/3}. \tag{6}$$

Plugging (6) into the $\log(4/\delta_0)$ term in (5), we have

$$\frac{1}{\delta_0} = \frac{C_1 K_{\alpha,\beta} \log(4/\delta_0)^{\beta'}}{\Delta_S^*} \leq \frac{(4/3)^{\beta'} C_1 K_{\alpha,\beta}}{\Delta_S^*} \cdot \log \left(\frac{4(8\beta')^{\beta'} C_1 K_{\alpha,\beta}}{\Delta_S^*} \right)^{\beta'}.$$

We define $D_{\alpha,\beta} = 4(8\beta')^{\beta'} K_{\alpha,\beta}$, $L_{\alpha,\beta} = (4/3)^{\beta'} K_{\alpha,\beta}$, then we have

$$\delta_0 \geq \frac{\Delta_S^*}{L_{\alpha,\beta} C_1 \log(D_{\alpha,\beta} C_1 / \Delta_S^*)^{\beta'}}.$$

By definition of i_0 we have that i_0 is the largest integer that satisfies $\gamma_{i_0} \geq \delta_0$. Therefore we get

$$\gamma_{i_0} \geq \frac{\Delta_S^*}{L_{\alpha,\beta} C_1 \log(D_{\alpha,\beta} C_1 / \Delta_S^*)^{\beta'}}. \tag{7}$$

Recall that $\gamma_i = 2^{-i(1-\alpha)/\alpha}$, (7) implies that

$$i_0 \leq \frac{\alpha}{1-\alpha} \log_2 \left(\frac{L_{\alpha,\beta} C_1 \log(D_{\alpha,\beta} C_1 / \Delta_S^*)^{\beta'}}{\Delta_S^*} \right),$$

or equivalently

$$2^{i_0} \leq \left(\frac{L_{\alpha,\beta} C_1 \log(D_{\alpha,\beta} C_1 / \Delta_S^*)^{\beta'}}{\Delta_S^*} \right)^{\frac{\alpha}{1-\alpha}}. \tag{8}$$

We divide the total satisficing regret into regret from rounds 1 to i_0 and after round i_0. Using the result of Proposition 1 we have that the satisficing regret by round i_0 is bounded by

$$\begin{aligned}
&\sum_{i=1}^{i_0} \frac{8C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^\beta \\
\leq &\sum_{i=1}^{i_0} \frac{8C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\delta_0)^\beta \\
\leq &\sum_{i=1}^{i_0} \frac{8C_1}{(1-\alpha)^\beta} 2^i \log(4/\delta_0)^\beta \\
\leq &\frac{32C_1}{(1-\alpha)^\beta} 2^{i_0} \log(4/\delta_0)^\beta \\
\leq &\frac{32(4/3)^\beta C_1}{(1-\alpha)^\beta} \left(\frac{L_{\alpha,\beta} C_1 \log(D_{\alpha,\beta} C_1 / \Delta_S^*)^{\beta'}}{\Delta_S^*} \right)^{\alpha/(1-\alpha)} \log \left(\frac{D_{\alpha,\beta} C_1}{\Delta_S^*} \right)^\beta \\
= &\frac{32(4/3)^\beta L_{\alpha,\beta}^{\alpha/(1-\alpha)} C_1^{1/(1-\alpha)}}{(1-\alpha)^\beta} \left(\frac{1}{\Delta_S^*} \right)^{\alpha/(1-\alpha)} \log \left(\frac{D_{\alpha,\beta} C_1}{\Delta_S^*} \right)^{\alpha\beta'/(1-\alpha)+\beta}.
\end{aligned} \tag{9}$$

Here the first inequality is because by definition of δ_0 and i_0, we have $\gamma_i = 2^{-i(1-\alpha)/\alpha} \geq \gamma_{i_0} \geq \delta_0$ for all $i \leq i_0$, the second inequality is by plugging into the definition of $\gamma_i = 2^{-i(1-\alpha)/\alpha}$, the third inequality is due to $\sum_{i=1}^{i_0} 2^i \leq 2^{i_0+1} \leq 4 \cdot 2^{i_0}$, and the fourth inequality is by plugging into (6) on $\log(4/\delta_0)$ and (8) on 2^{i_0}.

By the results of Proposition 2, for every $k \geq 1$, the probability that round $i_0 + k$ starts within the time horizon is at most 4^{-k+1}. Therefore the regret incurred after round i_0 is bounded by

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{8C_1}{(1-\alpha)^\beta} 2^{i_0+k} \log(4/\gamma_{i_0+k})^\beta \cdot 4^{-k+1} \\
\leq &\sum_{k=1}^{\infty} \frac{32C_1}{(1-\alpha)^\beta} 2^{i_0} \left(\log(4/\gamma_{i_0}) + \frac{1-\alpha}{\alpha} k \right)^\beta \cdot 2^{-k} \\
\leq &\sum_{k=1}^{\infty} \frac{32 \cdot 2^\beta C_1}{(1-\alpha)^\beta} 2^{i_0} \log(4/\gamma_{i_0})^\beta \cdot \frac{(1-\alpha)^\beta}{\alpha^\beta} k^\beta \cdot 2^{-k} \\
\leq &\frac{32 \cdot 2^\beta C_1}{\alpha^\beta} 2^{i_0} \log(4/\delta_0)^\beta \cdot \sum_{k=1}^{\infty} k^\beta \cdot 2^{-k} \\
\leq &\frac{32(8/3)^\beta L_{\alpha,\beta}^{\alpha/(1-\alpha)} C_1^{1/(1-\alpha)} M_\beta}{\alpha^\beta} \left(\frac{1}{\Delta_S^*} \right)^{\alpha/(1-\alpha)} \log \left(\frac{D_{\alpha,\beta} C_1}{\Delta_S^*} \right)^{\alpha\beta'/(1-\alpha)+\beta},
\end{aligned} \tag{10}$$

where $M_\beta = \sum_{k=1}^{\infty} k^\beta \cdot 2^{-k}$ is a constant only dependent on β. The first inequality follows directly from the definition of $\gamma_i = 2^{-i(1-\alpha)/\alpha} = \gamma_{i_0} \cdot 2^{-k(1-\alpha)/\alpha}$ with $i = i_0 + k$. The second inequality is because $\log(4/\gamma_{i_0}) > \log(4) > 1$ and $(1-\alpha)/\alpha \geq 1$ for $1/2 \leq \alpha < 1$, and for any $a, b \geq 1$, we have $2ab \geq ab + 1 \geq a + b$. The third inequality is because by definition $\delta_0 \leq \gamma_{i_0}$. The fourth inequality is due to (6) for $\log(4/\delta_0)$ and (8) for 2^{i_0}. Combining the satisficing regret before round i_0 (9) and after round i_0 (10), the total satisficing regret is bounded by

$$\mathbb{E}[\texttt{Regret}_S] \leq \frac{32(8/3)^\beta L_{\alpha,\beta}^{\alpha/(1-\alpha)} C_1^{1/(1-\alpha)} (M_\beta + 2^{-\beta})}{\alpha^\beta} \left(\frac{1}{\Delta_S^*} \right)^{\alpha/(1-\alpha)} \log \left(\frac{D_{\alpha,\beta} C_1}{\Delta_S^*} \right)^{\alpha\beta'/(1-\alpha)+\beta}. \tag{11}$$

Furthermore, since the length of the time horizon is T time steps, we have $t_i \leq T$. Since by definition of $t_i = \lceil 2^{i/\alpha} \rceil$, the number of rounds that end within the time horizon is bounded by $2^{i/\alpha} \leq T$, *i.e.*, $i \leq \lfloor \alpha \log_2(T) \rfloor$ because i is an integer. Then, the maximum number of rounds is bounded by $i_{\max} = \lceil \alpha \log_2(T) \rceil \leq \alpha \log_2(T) + 1$. Recall $\gamma_i = 2^{-i(1-\alpha)/\alpha}$, for any $i \leq i_{\max}$, we have

$$\begin{aligned}
\gamma_i \geq \gamma_{i_{\max}} &= 2^{-(1-\alpha)\lceil \alpha \log_2(T) \rceil/\alpha} \\
&\geq 2^{-(1-\alpha)\log_2(T)/\alpha - (1-\alpha)/\alpha} \\
&\geq 2^{-\log_2(T)-1} = 1/(2T),
\end{aligned} \tag{12}$$

where the third inequality is because $(1-\alpha)/\alpha \leq 1$ for $1/2 \leq \alpha < 1$. Therefore, we have that the satisficing regret is bounded by

$$\begin{aligned}
\mathbb{E}[\texttt{Regret}_S] \leq & \sum_{i=1}^{i_{\max}} \frac{8C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^\beta \\
\leq & \sum_{i=1}^{i_{\max}} \frac{8C_1}{(1-\alpha)^\beta} 2^{i+1} \log(8T)^\beta \\
\leq & \frac{32C_1}{(1-\alpha)^\beta} 2^{i_{\max}} \log(8T)^\beta \\
\leq & \frac{32C_1}{(1-\alpha)^\beta} 2^{\alpha \log_2(T)+1} \log(8T)^\beta \\
= & \frac{64C_1}{(1-\alpha)^\beta} T^\alpha \log(8T)^\beta,
\end{aligned} \tag{13}$$

where the first inequality is by the results of Proposition 1, the second inequality is by $4/\gamma_i \leq 8T$ (12) for all $i \leq i_{\max}$ and the definition $\gamma_i = 2^{-i(1-\alpha)/\alpha}$, the third inequality is by $\sum_{i=1}^{i_{\max}} 2^{i+1} \leq 2 \cdot 2^{i_{\max}+1} = 4 \cdot 2^{i_{\max}}$, and the fourth inequality is by $i_{\max} = \lceil \alpha \log_2(T) \rceil \leq \alpha \log_2(T) + 1$.

Combining the two parts of results, *i.e.*, the constant satisficing regret (11) and the sub-linear satisficing regret (13), we have completed the proof of Theorem 1 by putting all constant terms into the polylog term.

A.1. Proof of Proposition 1

We establish a satisficing regret bound for round i in three parts: Step 1 that runs the oracle $\texttt{ALG}(t_i)$, Step 2 that does forced sampling on the candidate arm, and Step 3 that pulls the candidate arm with LCB test.

Recall that $t_i = \lceil \gamma_i^{-1/(1-\alpha)} \rceil \leq 2\gamma_i^{-1/(1-\alpha)}$, then $\log(t_i) \leq \log(2\gamma_i^{-1/(1-\alpha)}) \leq \log(4/\gamma_i)/(1-\alpha)$ and $t_i^\alpha = \lceil 2^{i/\alpha} \rceil^\alpha \leq 2^{\alpha(1+i/\alpha)} \leq 2 \cdot 2^i = 2\gamma_i^{-\alpha/(1-\alpha)}$. Similarly, $t_i^{\alpha-1} \leq 2^{(\alpha-1)(1+i/\alpha)} = 2^{\alpha-1} \cdot 2^{i(\alpha-1)/\alpha} \leq 2\gamma_i$. With Condition 1 the satisficing regret incurred when running $\texttt{ALG}(t_i)$ is bounded by

$$\begin{aligned}
& \mathbb{E}\left[\sum_{s=1}^{t_i} \max\{S - r(A_s), 0\}\right] \\
\leq & \mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right] \leq C_1 t_i^\alpha \log(t_i)^\beta \\
\leq & \frac{2C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^\beta.
\end{aligned} \tag{14}$$

Since $\hat{A}_i$ is selected randomly according to the trajectory of $\texttt{ALG}(t_i)$ in round i, we further have

$$\begin{aligned}
&\mathbb{E}[\max\{S - r(\hat{A}_i), 0\}] \le \mathbb{E}[r(A^*) - r(\hat{A}_i)] \\
=&\frac{\mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right]}{t_i} \\
\le& C_1 t_i^{\alpha-1} \log(t_i)^\beta \\
\le& \frac{2C_1}{(1-\alpha)^\beta} \gamma_i \log(4/\gamma_i)^\beta.
\end{aligned} \tag{15}$$

Here the last inequality is because $\alpha < 1$ and $t_i = \lceil \gamma_i^{-1/(1-\alpha)} \rceil \ge \gamma_i^{-1/(1-\alpha)}$, therefore $t_i^{\alpha-1} \le \gamma_i$, and we also have $\log(t_i) \le \log(4/\gamma_i)/(1-\alpha)$ because $t_i = \lceil \gamma_i^{-1/(1-\alpha)} \rceil \le (2/\gamma_i)^{1/(1-\alpha)}$ as $1/(1-\alpha) \ge 2$. Therefore the satisficing regret incurred when pulling $\hat{A}_i$ for $T_i = \lceil \gamma_i^{-2} \rceil$ times during forced sampling is bounded by

$$\begin{aligned}
&\left\lceil \frac{1}{\gamma_i^2} \right\rceil \cdot \frac{2C_1}{(1-\alpha)^\beta} \gamma_i \log(4/\gamma_i)^\beta \\
\le& \frac{4C_1}{(1-\alpha)^\beta} \gamma_i^{-1} \log(4/\gamma_i)^\beta \\
\le& \frac{4C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^\beta,
\end{aligned} \tag{16}$$

where the first inequality is because $1/\gamma_i^2 \ge 1$ and then $\lceil \gamma_i^{-2} \rceil \le 2\gamma_i^{-2}$, the second inequality is because $\alpha/(1-\alpha) \ge 1$ for $0.5 \le \alpha < 1$.

For the LCB test, we only consider the case where the candidate arm $\hat{A}_i$ is in fact non-satisficing, *i.e.*, $r(\hat{A}_i) < S$, because pulling satisficing ones incurs zero satisficing regret. Then for every k, round i is not terminated after pulling non-satisficing arm $\hat{A}_i$ for k times (in addition to the first T_i times) occurs with probability at most

$$\begin{aligned}
&\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i + k} - \sqrt{\frac{4\log(T_i + k)}{T_i + k}} > S \right) \\
\le& \Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i + k} - \sqrt{\frac{4\log(T_i + k)}{T_i + k}} > r(\hat{A}_i) \right) \\
\le& \exp\left(-2 \cdot 4\log(T_i + k)\right) \\
=& \frac{1}{(T_i + k)^8} \le \frac{1}{(k+1)^8},
\end{aligned}$$

where the second inequality is due to Hoeffding's inequality, and the third inequality due to $T_i \ge 1$. Therefore expected number of arm pulls of $\hat{A}_i$ after first pulling it T_i times is bounded by

$$\begin{aligned}
&\mathbb{E}\left[\text{Number of times } \hat{A}_i \text{ is pulled in round } i \text{ after the first } T_i \text{ times}\right] \\
=& \sum_{s=1}^{\infty} \Pr(k \ge s) = \sum_{s=1}^{\infty} \frac{1}{(s+1)^8} \le 1.
\end{aligned}$$

Since we have assumed that the expected reward is bounded on $[0,1]$ for every arm, the expected satisficing regret incurred in step 3 is upper bounded by 1.

Summing up the three parts of satisficing regret, *i.e.*, the satisficing regret from Step 1 (14), Step 2 (16), and Step 3 (combination of (17) and (15)), we have that the satisficing regret of round i is bounded by

$$\begin{aligned}&\frac{2C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta+\frac{4C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta+1\cdot\frac{2C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^\beta\\ \leq&\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.\end{aligned}\tag{17}$$

A.2. Proof of Proposition 2

We denote $A^*=\arg\max_{A\in\mathcal{A}}r(A)$.

By Condition 1 and recall from (15) in the proof of Proposition 1, we have

$$\mathbb{E}[r(A^*)-r(\hat{A}_i)]\leq\frac{2C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^\beta.\tag{18}$$

According to the condition of this proposition 2, we have i satisfies the condition

$$\frac{32C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^{\max\{\beta,1/2\}}\leq\Delta_S^*.$$

By the Markov's inequality we have

$$\Pr(r(A^*)-r(\hat{A}_i)\geq\Delta_S^*/2)\leq\Pr\left(r(A^*)-r(\hat{A}_i)\geq\frac{16C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^\beta\right)\leq\frac{1}{8},\tag{19}$$

where we have $\log(4/\gamma_i)\geq\log(4)>1$ given $\gamma_i<1$ for any i.

If $r(\hat{A}_i)\geq r(A^*)-\Delta_S^*/2$, which happens with probability at least $7/8$ according to the above inequality, then $r(\hat{A}_i)\geq S+\Delta_S^*/2$ holds. On the other hand, if i satisfies the condition stated in the proposition, then since $C_1\geq1$ and $(1-\alpha)^\beta<1$ according to the assumption in Condition 1, we have $\gamma_i\log(4/\gamma_i)^{1/2}\leq\Delta_S^*/32$. Plugging into the confidence radius, we have

$$\begin{aligned}\sqrt{\frac{4\log(T_i+k)}{T_i+k}}\leq&4\sqrt{\frac{\log(T_i)}{T_i}}=4\sqrt{\frac{\log(\lceil\gamma_i^{-2}\rceil)}{\lceil\gamma_i^{-2}\rceil}}\\ \leq&4\sqrt{\gamma_i^2\log(2\gamma_i^{-2})}\leq4\sqrt{2}\gamma_i\log(4/\gamma_i)^{1/2}\\ \leq&4\sqrt{2}\cdot\frac{\Delta_S^*}{32}\leq\frac{\Delta_S^*}{4},\end{aligned}\tag{20}$$

where the first inequality is because $T_i>1$, and $f(y)=\log(y)/y$ is monotonically decreasing when $y\geq3$ and $2\cdot\log(2)/2\geq\log(3)/3$ for all $k\geq0$, thus $\log(T_i+k)/(T_i+k)\leq4\log(T_i)/T_i$ for all $k\geq0$; the second inequality is because $\gamma_i^{-1}\geq1$, thus $\lceil\gamma_i^{-2}\rceil\leq2\gamma_i^{-2}$. Suppose round i is terminated for some k, then

$$\frac{\hat{r}_i^{\text{tot}}}{T_i+k}-\sqrt{\frac{4\log(T_i+k)}{T_i+k}}<S=r(A^*)-\Delta_S^*\leq r(\hat{A}_i)-\frac{\Delta_S^*}{2}\leq r(\hat{A}_i)-2\sqrt{\frac{4\log(T_i+k)}{T_i+k}},$$

or equivalently

$$\frac{\hat{r}_i^{\text{tot}}}{T_i+k} \le r(\hat{A}_i) - \sqrt{\frac{4\log(T_i+k)}{T_i+k}}.$$

By the Hoeffding's inequality we have

$$\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i+k} \le r(\hat{A}_i) - \sqrt{\frac{4\log(T_i+k)}{T_i+k}}\right) \le \frac{1}{(T_i+k)^8} \le \frac{1}{(k+1)^8}.$$

Therefore conditioned on $r(\hat{A}_i) \ge r(A^*) - \Delta_S^*/2$, the probability that round i is terminated by the end of the horizon is bounded by

$$\sum_{k=1}^{\infty} \frac{1}{(k+1)^8} \le \frac{1}{8}. \tag{21}$$

Combining (19) and (21) we conclude that the probability that round i terminates by the end of the time horizon is upper bounded by $1/4$.

B. Proof of Theorem 2

We first prove that the standard regret incurred in round i is at most

$$\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

The regret incurred when running $\texttt{ALG}(t_i)$ in Step 1 is bounded by

$$\mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right] \le C_1 t_i^\alpha \log(t_i)^\beta \le \frac{2C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

Recall that in (18) we have shown that

$$\mathbb{E}[r(A^*) - r(\hat{A}_i)] \le \frac{2C_1}{(1-\alpha)^\beta}\gamma_i \log(4/\gamma_i)^\beta.$$

Therefore similar to (16), the standard regret incurred when pulling $\hat{A}_i$ in Step 2 forced sampling is bounded by

$$\left\lceil \frac{1}{\gamma_i^2} \right\rceil \cdot \frac{2C_1}{(1-\alpha)^\beta}\gamma_i \log(4/\gamma_i)^\beta \le \frac{4C_1}{(1-\alpha)^\beta}\gamma_i^{-1}\log(4/\gamma_i)^\beta \le \frac{4C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

Since $r(\hat{A}_i) \le r(A^*) < S$, similar to (17), for every k, round i is not terminated with probability at most

$$\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i+k} - \sqrt{\frac{4\log(T_i+k)}{T_i+k}} > S\right) \le \frac{1}{(k+1)^8}.$$

Therefore expected number of arm pulls of $\hat{A}_i$ after pulling T_i times in Step 3 is bounded by

$$\sum_{k=1}^{\infty} \frac{1}{(k+1)^8} \le 1.$$

Summing up the standard regret from Step 1, 2, and 3 in a round i, we have that the standard regret of round i is bounded by

$$\begin{aligned}&\frac{2C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta+\frac{4C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta+1\cdot\frac{2C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^\beta\\ \le&\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.\end{aligned}$$

Furthermore, since the length of the time horizon is T time steps, we have $t_i \le T$. Since by definition of $t_i = \lceil 2^{i/\alpha} \rceil$, the number of rounds that end within the time horizon is bounded by $2^{i/\alpha} \le T$, *i.e.*, $i \le \lfloor \alpha \log_2(T) \rfloor$ because i is an integer. Then, the maximum number of rounds is bounded by $i_{\max} = \lceil \alpha \log_2(T) \rceil$. Following the same chain of inequalities in (13), we can bound the standard regret by

$$\begin{aligned}\mathbb{E}[\texttt{Regret}] \le& \sum_{i=1}^{i_{\max}} \frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta\\ \le& \frac{64C_1}{(1-\alpha)^\beta}T^\alpha \log(4T)^\beta.\end{aligned}$$

Therefore, we have proven that the regret is bounded by $C_1 T^\alpha \cdot \text{polylog}(T)$.

C. Proofs for Lower Bounds

C.1. Proof of Theorem 3

We consider the following K instance of finite-arm bandits with K arms. We set the satisficing threshold $S = 1/2 + \Delta$. The mean reward of each arm follows a Bernoulli distribution. For each $k \in [K]$, the mean rewards of the k-th instance is defined as

$$r_k = \frac{1}{2} + 2\Delta,\ r_{k'} = \frac{1}{2},\ \forall k \ne k'.$$

Similar lower bound instances are also considered in Section 2 of Slivkins et al. (2019). By Lemma 2.8 of Slivkins et al. (2019), there exists $0 < c \le 1$ such that for any policy that aims to identify the optimal arm within cK/Δ^2 time steps, there exists $\lceil K/3 \rceil$ instances such that the policy incorrectly identify the optimal arm with probability at least $1/4$. We formally restate the lemma in Lemma 1 in Section P of the appendix. If we consider the arm pull of policy π at time step t as a policy that commits one arm A_t after $t-1$ arm pulls, then by Lemma 1 we have that for all $t \le cK/\Delta^2$, there exists at least $\lceil K/3 \rceil$ arms $k \in [K]$ such that $\Pr^{(k)}(A_t = k) \le 3/4$. Therefore we have that $\sum_{k=1}^K \Pr^{(k)}(A_t \ne k) \ge K/3 \cdot 1/4 = K/12$. Since under any instance $k \in [K]$, pulling any arm $k' \ne k$ incurs a satisficing regret of Δ. Therefore we have

$$\sum_{k=1}^K \mathbb{E}^{(k)}[\texttt{Regret}_S] = \sum_{t=1}^T \sum_{k=1}^K \Delta \Pr^{(k)}(A_k \ne k) \ge \sum_{t=1}^{\lfloor cK/\Delta^2 \rfloor} \sum_{k=1}^K \Delta \Pr^{(k)}(A_k \ne k) \ge \frac{\Delta K}{12} \cdot \left\lfloor \frac{cK}{\Delta^2} \right\rfloor = \Omega\left(\frac{K^2}{\Delta}\right).$$

Here the first inequality is due to the assumption that $T \ge K/\Delta^2$ and $c \le 1$. Therefore there exists $k \in [K]$ such that the expected satisficing regret $\mathbb{E}^{(k)}[\texttt{Regret}_S]$ is at least $\Omega(K/\Delta)$.

C.2. Proof of Theorem 4

We define the arm set as $\mathcal{A}=[-1,1]$. We consider the following two instances of concave bandits. The reward functions of the two instances are defined as follows.

$$\textbf{Instance 1: } r^{(1)}(x)=\begin{cases}2\sqrt{2\Delta}x+6\Delta+\frac{1}{2}, & \text{if } -1\le x\le -2\sqrt{2\Delta},\\ \frac{1}{2}-(x+\sqrt{2\Delta})^2, & \text{if } -2\sqrt{2\Delta}<x<0\\ -2\sqrt{2\Delta}x+\frac{1}{2}-2\Delta, & \text{if } 0\le x\le 1;\end{cases}$$

$$\textbf{Instance 2: } r^{(2)}(x)=\begin{cases}2\sqrt{2\Delta}x+\frac{1}{2}-2\Delta, & \text{if } -1\le x\le 0,\\ \frac{1}{2}-(x-\sqrt{2\Delta})^2, & \text{if } 0<x<2\sqrt{2\Delta},\\ -2\sqrt{2\Delta}x+6\Delta+\frac{1}{2}, & \text{if } 2\sqrt{2\Delta}\le x\le 1.\end{cases}$$

Since we have $\Delta<1/162$, we have $1/2\ge r^{(1)}(x)\ge 1/2-2\Delta-2\sqrt{2\Delta}\ge 1/2-1/81-2/9>1/4$ for all $x\in[-1,1]$. Similarly, we also have $1/2\ge r^{(2)}(x)\ge 1/4$ for all $x\in[-1,1]$.

We set $S=1/2-\Delta$, and the random reward of arm x follows a Bernoulli distribution with mean $r^{(i)}(x)$ under instance i for $i\in\{1,2\}$. Since we have proven that $r^{(1)}(x),r^{(2)}(x)\in[0,1]$ for all $x\in[-1,1]$, the definition of random reward is valid. Since for both instances, the optimal reward is $1/2$, thus both instances satisfies $r(A^*)-S=\Delta$.

By definition of $r^{(1)}$ and $r^{(2)}$ we have

$$\frac{\mathrm{d}r^{(1)}}{\mathrm{d}x}(x)=\begin{cases}2\sqrt{2\Delta}, & \text{if } x<-2\sqrt{2\Delta},\\ -2x-2\sqrt{2\Delta}, & \text{if } -2\sqrt{2\Delta}\le x\le 0,\\ -2\sqrt{2\Delta}, & \text{if } x>0,\end{cases}$$

$$\frac{\mathrm{d}r^{(2)}}{\mathrm{d}x}(x)=\begin{cases}2\sqrt{2\Delta}, & \text{if } x<0,\\ -2x+2\sqrt{2\Delta}, & \text{if } 0\le x\le 2\sqrt{2\Delta},\\ -2\sqrt{2\Delta}, & \text{if } x>2\sqrt{2\Delta}.\end{cases}$$

Therefore we have $r^{(1)}(x)-r^{(2)}(x)$ is monotonically non-increasing, thus $r^{(1)}(x)-r^{(2)}(x)$ reaches maximum when $x=-1$ and reaches minimum when $x=1$. We further get

$$r^{(1)}(-1)-r^{(2)}(-1)=8\Delta,\ r^{(1)}(1)-r^{(2)}(1)=-8\Delta.$$

Therefore we have that $|r^{(1)}(x)-r^{(2)}(x)|\le 8\Delta$ for all $x\in[-1,1]$.

Next, we prove that for any $p,q\in[1/4,1/2]$, we have $\mathtt{KL}(\text{Bernoulli}(p)||\text{Bernoulli}(q))\le 8(p-q)^2$. For any $p\in[1/4,1/2]$, we define $f_p(q)$ as

$$f_p(q)=\mathtt{KL}(\text{Bernoulli}(p)||\text{Bernoulli}(q))=p\log\left(\frac{p}{q}\right)+(1-p)\log\left(\frac{1-p}{1-q}\right).$$

Then we have $f_p(p)=0$ and $f_p'(q)=-p/q+(1-p)/(1-q)$, thus $f_p'(p)=0$. We further have that for all $p,q\in[1/4,1/2]$,

$$f_p''(q)=\frac{p}{q^2}+\frac{1-p}{(1-q)^2}\le\frac{p}{(1/4)^2}+\frac{1-p}{(1/2)^2}=4+12p\le 10.$$

Here the first inequality is due to $q \in [1/4, 1/2]$ and thus $1-q \in [1/2, 3/4]$. Furthermore, by Taylor's theorem, there exists $\xi \in (p, q)$ such that

$$f_p(q) = f_p(p) + f_p'(p)(q-p) + \frac{1}{2} f_p''(\xi)(q-p)^2.$$

Since $f_p(p) = f_p'(p) = 0$, we have $f_p(q) = f_p''(\xi)(q-p)^2/2 \le 5(q-p)^2 \le 8(q-p)^2$, therefore for all $p, q \in [1/4, 1/2]$, we have $\mathtt{KL}(\text{Bernoulli}(p)||\text{Bernoulli}(q)) \le 8(q-p)^2$. Since we have proven that $|r^{(1)}(x) - r^{(2)}(x)| \le 8\Delta$, we have

$$\mathtt{KL}\left(\text{Bernoulli}(r^{(1)}(x))\middle|\middle|\text{Bernoulli}(r^{(2)}(x))\right) \le 8(r^{(1)}(x) - r^{(2)}(x))^2 \le 512\Delta^2.$$

We define $T_0 = \lfloor 1/512\Delta^2 \rfloor$, and define $\Pr_t^{(i)}$ as the probability space of a policy π by time step t under instance i. Then we have

$$\mathtt{KL}(\Pr_t^{(1)}||\Pr_t^{(2)}) \le t \cdot \max_{x \in [-1,1]} \mathtt{KL}\left(\text{Bernoulli}(r^{(1)}(x))\middle|\middle|\text{Bernoulli}(r^{(2)}(x))\right) \le t \cdot 512\Delta^2 \le T_0 \cdot 512\Delta^2 \le 1. \tag{22}$$

Here the second inequality follows from (22). By the Bretagnolle-Huber inequality (see, *e.g.*, Theorem 14.2 of Lattimore and Szepesvári 2020), we have that for all $t \le T_0$,

$$\Pr^{(1)}(A_t > 0) + \Pr^{(2)}(A_t \le 0) \ge \frac{1}{2}\exp(-\mathtt{KL}(\Pr_t^{(1)}||\Pr_t^{(2)})) \ge \frac{1}{2}\exp(-1). \tag{23}$$

We denote $\mathtt{Regret}_S^{(i)}$ as the satisficing regret under instance i. By definition of $r^{(1)}$ and $r^{(2)}$ we have $r^{(1)}(x) \le 1 - 2\Delta$ for all $x > 0$ and $r^{(2)}(x) \le 1 - 2\Delta$ for all $x < 0$. Since $S = 1 - \Delta$, pulling any arm $x > 0$ under instance 1 and pulling any arm $x < 0$ under instance 2 incurs a satisficing regret of at least Δ. Therefore we have

$$\begin{aligned}
\mathbb{E}^{(1)}[\mathtt{Regret}_S] + \mathbb{E}^{(2)}[\mathtt{Regret}_S] &\ge \sum_{t=1}^{T} \Delta(\Pr^{(1)}(A_t > 0) + \Pr^{(2)}(A_t < 0)) \\
&\ge \sum_{t=1}^{T_0} \Delta(\Pr^{(1)}(A_t > 0) + \Pr^{(2)}(A_t < 0)) \\
&\ge T_0 \Delta \cdot \frac{1}{2}\exp(-1) = \frac{1}{2}\exp(-1)\Delta\left\lfloor \frac{1}{512\Delta^2} \right\rfloor = \Omega\left(\frac{1}{\Delta}\right).
\end{aligned}$$

Here the first inequality is due to the assumption that $T \ge 1/\Delta^2$, and the second inequality is follows from (23). Therefore under either instance 1 or instance 2, the policy incurs $\Omega(1/\Delta)$ regret.

C.3. Proof of Theorem 5

We define $M = \lfloor L/4\Delta \rfloor$, and we define M^d instances indexed by $\boldsymbol{m} \in [M]^d$. The arm set is defined as $[0,1]^d$, and the reward of each arm follows a Bernoulli distribution. For every $\boldsymbol{m} \in [M]^d$, we define $\boldsymbol{y}^{(\boldsymbol{m})}$ as $y_j^{\boldsymbol{m}} = (2m_j - 1)/2M$, and the mean reward function $r^{(\boldsymbol{m})}$ of instance $\boldsymbol{m}$ is defined as

$$r^{(\boldsymbol{m})}(\boldsymbol{x}) = \max\left\{\frac{1}{2}, \frac{1}{2} + 2\Delta - L||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty\right\}.$$

For any $\boldsymbol{x} \in [0,1]^d$ and $\boldsymbol{m} \in [M]^d$ such that $||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty \geq 1/2M$, since $M = \lfloor L/4\Delta \rfloor \leq L/4\Delta$, we have $||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty \geq 2\Delta/L$, thus $r^{(\boldsymbol{m})}(\boldsymbol{x}) = 1/2$. Therefore $r^{(\boldsymbol{x})}(\boldsymbol{x}) > 1/2$ only when $||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty \leq 1/2M$.

We also construct M^d instances of finite armed bandit indexed by $[M]^d$. The reward function of instance $\boldsymbol{m}$ is denoted as $\bar{r}^{(\boldsymbol{m})}$. The arm set is set to be $[M]^d$, and for each $\boldsymbol{m} \in [M]^d$, we define instance $\boldsymbol{m}$ as

$$\bar{r}^{(\boldsymbol{m})}(\boldsymbol{m}) = \frac{1}{2} + 2\Delta, \ \bar{r}^{(\boldsymbol{m})}(\boldsymbol{m}') = \frac{1}{2}, \ \forall \boldsymbol{m}' \neq \boldsymbol{m}. \tag{24}$$

Based on the policy π, we construct a policy π' for finite-arm bandits with arm set being $[M]^d$. For any $\boldsymbol{x} \in [0,1]^d$, we define $\boldsymbol{m}(\boldsymbol{x}) = \arg\min\{||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}|| : \boldsymbol{m} \in [M]^d\}$. For each round $t \in [T]$, suppose the policy π pulls arm $\boldsymbol{m}(\boldsymbol{x})$ and observes a random reward R. Then, with probability $p_{\boldsymbol{x}}$, we feed the random reward R back to policy π, and with probability $1 - p_{\boldsymbol{x}}$, we generate a Bernoulli random variable with mean $1/2$ and feed the realized random variable back to policy π. Here, we define

$$p_{\boldsymbol{x}} = \frac{1}{2\Delta} \max\{2\Delta - L||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m}(\boldsymbol{x}))}||_\infty, 0\}.$$

We show that under any instance $\boldsymbol{m} \in [M]^d$, the random reward fed back to π follows a Bernoulli distribution of $r^{(\boldsymbol{m})}(\boldsymbol{x})$. For any $\boldsymbol{m} \neq \boldsymbol{m}(\boldsymbol{x})$, the random reward R of arm $\boldsymbol{m}$ follows a Bernoulli distribution with mean $1/2$, thus the reward fed back to π also follows a distribution with mean $1/2$, which is equal to $r^{(\boldsymbol{m})}(\boldsymbol{x})$ since $\boldsymbol{m} \neq \boldsymbol{m}(\boldsymbol{x})$. If $\boldsymbol{m} = \boldsymbol{m}(\boldsymbol{x})$, then the probability the reward fed back to π is equal to 1 is equal to

$$\begin{aligned}
&p_{\boldsymbol{x}} \bar{r}^{(\boldsymbol{m}(\boldsymbol{x}))}(\boldsymbol{m}(\boldsymbol{x})) + (1 - p_{\boldsymbol{x}}) \cdot \frac{1}{2} = p_{\boldsymbol{x}} \left(\frac{1}{2} + 2\Delta \right) + (1 - p_{\boldsymbol{x}}) \cdot \frac{1}{2} \\
=&\frac{1}{2} + 2\Delta p_{\boldsymbol{x}} = \frac{1}{2} + \max\{2\Delta - L||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m}(\boldsymbol{x}))}||_\infty, 0\} \\
=&\max\left\{ \frac{1}{2} + 2\Delta - L||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m}(\boldsymbol{x}))}||_\infty, \frac{1}{2} \right\} = r^{(\boldsymbol{m}(\boldsymbol{x}))}(\boldsymbol{x}).
\end{aligned}$$

Therefore under any instance $\boldsymbol{m} \in [M]^d$, the mean reward fed back to π is equal to $r^{(\boldsymbol{m})}(\boldsymbol{x})$, thus the reward observed by the policy π in this setting follows the same distribution as running the policy on the corresponding Lipschitz bandit instance.

Next, we show that the expected satisficing regret incurred by π is greater than or equal to that incurred by π'. It suffices to prove that $r^{(\boldsymbol{m})}(\boldsymbol{x}) \leq \bar{r}^{(\boldsymbol{m})}(\boldsymbol{m}(\boldsymbol{x}))$. If $\boldsymbol{m} = \boldsymbol{m}(\boldsymbol{x})$, then by definition we have $\bar{r}^{(\boldsymbol{m})}(\boldsymbol{m}(\boldsymbol{x})) = 1/2 + 2\Delta \geq r^{(\boldsymbol{m})}(\boldsymbol{x})$. If $\boldsymbol{m}(\boldsymbol{x}) \neq \boldsymbol{m}$, we define $\boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{m}'$. By definition of $\{\boldsymbol{y}^{(\boldsymbol{m})} : \boldsymbol{m} \in [M]^d\}$, for every $\boldsymbol{x} \in [0,1]^d$, there exists $\boldsymbol{m} \in [M]^d$ such that $||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty \leq 1/M$. Since by definition $\boldsymbol{y}^{(\boldsymbol{m}')}$ is the closest to $\boldsymbol{x}$ among $\{\boldsymbol{y}^{(\boldsymbol{m})} : \boldsymbol{m} \in [M]^d\}$ measured by infinite norm, we have $||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m}')}|| \leq 1/M$. Furthermore, since $\boldsymbol{m} \neq \boldsymbol{m}'$, by definition of $\boldsymbol{y}^{(\boldsymbol{m})}$ we have $||\boldsymbol{y}^{(\boldsymbol{m})} - \boldsymbol{y}^{(\boldsymbol{m}')}||_\infty \geq 1/M$, therefore

$$||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty \geq ||\boldsymbol{y}^{(\boldsymbol{m})} - \boldsymbol{y}^{(\boldsymbol{m}')}||_\infty - ||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty \geq \frac{1}{M} - \frac{1}{2M} = \frac{1}{2M}.$$

We have proven that $r^{(\boldsymbol{m})}(\boldsymbol{x}) = 1/2$ for all $\boldsymbol{x} \in [0,1]^d$ such that $||\boldsymbol{x} - \boldsymbol{y}^{(\boldsymbol{m})}||_\infty \geq 1/2M$. Therefore in this case we also have $\bar{r}^{(\boldsymbol{m})}(\boldsymbol{m}(\boldsymbol{x})) \geq 1/2 = r^{(\boldsymbol{m})}(\boldsymbol{x})$. Then we conclude that the mean reward of any arm pull of π is smaller than or equal to that of π', thus under any instance $\boldsymbol{m} \in [M]^d$, the expected satisficing regret of π is greater than or equal to that of π'.

By Theorem 3, when $T \geq M^d/\Delta^2$, there exists an instance $\boldsymbol{m} \in [M]^d$ as constructed in (24) such that the expected satisficing regret of π' is at least $\Omega(M^d/\Delta)$. Since $M = \Theta(L/\Delta)$, and the expected satisficing regret of π is greater than or equal to that of π', we have that there exists an instance such that the expected satisficing regret of π is at least $\Omega(L^d/\Delta^{d+1})$.

D. Proof of Proposition 3

We consider a simple instance of two-armed bandit: $r(1) = 1$ and $r(2) = 0$ with $S = 0.5$.

Since for finite-armed bandit, there exists algorithms that attains a $\tilde{O}(\sqrt{T})$ regret bound, we have $\alpha = 1/2$, thus by definition of $t_i = \lceil 2^{i/\alpha} \rceil$. For ease of discussion, we assume that the algorithm `SELECT`runs indefinitely, but only the satisficing regret incurred during the first T time periods is considered. We define $i_{\max} = \lceil \alpha \log_2(T) \rceil$, and define $\mathcal{E}$ as the event that in the infinite time horizon, there are at least $i_{\max}$ rounds that are terminated. Recall that we have assumed that the probability that arm 2 is pulled at least once in each round $i \in [l]$ is at least 1/2. Define event $\mathcal{E} := \{\text{arm 2 is pulled in Step 2 for every round } i \in [l]\}$, and $\mathcal{E}_i := \{\text{arm 2 is pulled in Step 2 in round } i\}$. Then for each round i, $\mathcal{E}_i$ happens if arm 2 is pulled at least once in Step 1, and is further sampled to Step 2. According to the assumption in Proposition 3, the probability that arm 2 is pulled at least once in Step 1 is at least 1/2. During the random sampling that proposes a candidate arm at the end of Step 1, arm 2 is sampled as candidate with probability at least $1/t_i$ given it is pulled at least once in Step 1. Then, $\Pr(\mathcal{E}_i) \geq 1/2 \cdot 1/t_i$ for any i, $\mathcal{E}$ happens with probability at least

$$
\begin{aligned}
\Pr(\mathcal{E}) &= \prod_{i=1}^{l} \Pr(\mathcal{E}_i) \geq \prod_{i=1}^{l} \frac{1}{2} \cdot \frac{1}{t_i} \geq \prod_{i=1}^{i_{\max}} \frac{1}{2} \cdot \frac{1}{t_i} \\
&\overset{\text{(i)}}{\geq} 2^{-(1+\log_2(T)/2)} \cdot \prod_{i=1}^{1+\log_2(T)/2} 2^{-(1+2i)} = 2^{-(2+\log_2(T))} \cdot \prod_{i=1}^{1+\log_2(T)/2} 2^{-2i} \\
&= \frac{1}{4T} \cdot \prod_{i=1}^{1+\log_2(T)/2} 2^{-2i} = \frac{1}{4T} \cdot 2^{-(2+\log_2(T)/2)(1+\log_2(T)/2)} \\
&\overset{\text{(ii)}}{\geq} \frac{1}{4T} \cdot 2^{-\log_2(T)^2} = \frac{1}{4} \cdot \frac{\exp(-\log(2)\log_2(T)^2)}{\exp(\log(T))} \\
&= \frac{1}{4} \exp(-(\log(T)^2/\log(2) + \log(T))) \\
&\overset{\text{(iii)}}{\geq} \frac{1}{4} \exp(-2\log(T)^2).
\end{aligned} \tag{25}
$$

where step (i) is based on $i_{\max} \le 1 + \log_2(T)/2$ and the definition of $t_i = \lceil 2^{i/\alpha} \rceil \le 2^{2i+1}$ for $\alpha = 1/2$, step (ii) holds for $T \ge 8$, and step (iii) holds with $1 + \log(T)/\log(2) \le 2\log(T)$ for $T \ge 8$.

Next, we show that if $\mathcal{E}$ holds, then the satisficing regret of `SELECT` is at least $T/14$. By definition we have $t_{i_{\max}} = 2^{2i_{\max}} \ge T$, therefore within a time horizon of T time steps, the total number of rounds cannot exceed $i_{\max}$. Therefore if $\mathcal{E}$ holds, then for each round i that is started within a time horizon of T time periods, the candidate satisficing arm $\hat{A}_i$ is arm 2. If the total number of time steps for Step 3 is at least $T/7$, the total satisficing regret is at least $T/7 \cdot 1/2 = T/14$. Otherwise, the total number of time steps for Step 1 and 2 is at least $6T/7$. In this case, we claim that the total number of time steps for Step 2 is at least $T/7$. Let i_0 be the number of rounds that terminates within T time steps. Then the total number of time steps for Steps 2 is at least $\sum_{i=1}^{i_0} 4^i$, and the total number of time steps for Step 1 is at most $\sum_{i=1}^{i_0+1} 4^i$. The ratio between the two is at least

$$\frac{\sum_{i=1}^{i_0} 4^i}{\sum_{i=1}^{i_0+1} 4^i} \le \frac{4^{i_0}}{4^{i_0} + 4^{i_0+1}} = \frac{1}{5}.$$

Therefore the total number of time steps for Step 1 is at most 5 times that of Step 2, thus the total number of time steps for Step 1 is at least $T/7$. Since under $\mathcal{E}$, `SELECT` always pulls arm 2 in Step 2, the satisficing regret incurred during Step 2 is at least $T/7 \cdot 1/2 = T/14$. Combining both cases, we know that the satisficing regret is always at least $T/14$ under $\mathcal{E}$. Therefore, we can bound the probability of total satisficing regret exceeding $T/14$ as

$$\Pr\left(\texttt{Regret}_S \ge \frac{T}{14}\right) \ge \Pr(\mathcal{E}) \ge \frac{1}{4}\exp(-2\log(T)^2),$$

where the final inequality is based on (25).

E. Proof of Theorem 6

The inequality in the theorem is trivial when $x < x_0$ as the right-hand side is at least $\exp(2i_0 + 4) > 1 \ge \Pr(\texttt{Regret}_S > x)$, thus for the rest of the proof we assume $x \ge x_0$.

Denote $\beta' = \max\{\beta, 1/2\}$ and $K_{\alpha,\beta,\zeta} = \frac{32C_1(1-\zeta)^{-\zeta/2}}{(1-\alpha)^\beta}\sqrt{2 + 2\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}$, and i_0 as

$$i_0 = \max\{i : K_{\alpha,\beta,\zeta}\gamma_i^{1-\zeta}\log(4/\gamma_i)^{\max\{\beta,1/2\}} > \Delta_S^*\}.$$

Define the satisficing regret incurred in the Step 1 and 2 in each round i as R_i, then the probability that cumulative satisficing regret incurred in the Step 1 and 2 after round i_0 exceeds $x/3$ is upper bounded by

$$\begin{aligned}
\Pr\left(\sum_{i=i_0+1}^{\infty} R_i > \frac{x}{3}\right) &= \sum_{k=1}^{\infty}\Pr\left(\sum_{i=i_0+1}^{\infty} R_i > \frac{x}{3} \text{ and algorithm ends at round } i_0 + k\right) \\
&= \sum_{k=1}^{\infty}\Pr\left(\sum_{i=i_0+1}^{\infty} R_i > \frac{x}{3}\middle|\text{Ends at round } i_0 + k\right)\cdot\Pr(\text{Ends at round } i_0 + k) \\
&\le \sum_{k=1}^{\infty}\Pr\left(\sum_{i=i_0+1}^{i_0+k} R_i > \frac{x}{3}\right)\cdot 4^{1-k},
\end{aligned}$$

where the final inequality is based on Proposition 4.

We first prove an upper bound on i_0. Denote $\delta_0 \in (0,1)$ the smallest solution to the following equation

$$\delta_0^{1-\zeta} \log(4/\delta_0)^{\beta'} = \frac{\Delta_S^*}{K_{\alpha,\beta,\zeta}}. \tag{26}$$

Here δ_0 always exists because the following observations: $\delta^{1-\zeta}\log(4/\delta)^{\beta'}$ is continuous on $(0,1]$, $\delta^{1-\zeta}\log(4/\delta)^{\beta'} \to 0$ when $\delta \to 0$, and $\delta^{1-\zeta}\log(4/\delta)^{\beta'} > 1$ when $\delta = 1$; furthermore, $\Delta_S^* \le 1$ and $K_{\alpha,\beta,\zeta} \ge 1$, thus the right hand side lies in $(0,1]$. Therefore such δ_0 always exists. Then we have

$$\begin{aligned}
\frac{1}{\delta_0^{1-\zeta}} =& \frac{K_{\alpha,\beta,\zeta}\log(4/\delta_0)^{\beta'}}{\Delta_S^*} \\
=& \frac{K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'} \cdot (4\beta')^{\beta'} \log\left((4/\delta_0)^{(1-\zeta)/(4\beta')}\right)^{\beta'}}{\Delta_S^*} \\
\le& \frac{K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'} \cdot (4\beta')^{\beta'} \left(4/\delta_0^{1-\zeta}\right)^{1/4}}{\Delta_S^*} \\
=& \frac{K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'} \cdot (4\beta')^{\beta'} 2^{1/2}}{\Delta_S^* \left(\delta_0^{1-\zeta}\right)^{1/4}} \\
\le& \frac{(8\beta')^{\beta'} K_{\alpha,\beta}/(1-\zeta)^{\beta'}}{\Delta_S^* \left(\delta_0^{1-\zeta}\right)^{1/4}},
\end{aligned}$$

where the first inequality is because $\log(x) < x$ for all $x > 0$ and $2^{1-\zeta} \le 2$, and the second inequality is because $\beta' = \max\{\beta, 1/2\} \ge 1/2$, thus $2^{1/2} \le 2^{\beta'}$. Therefore

$$\frac{1}{\delta_0^{1-\zeta}} \le \left(\frac{(8\beta')^{\beta'} K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'}}{\Delta_S^*}\right)^{4/3}, \tag{27}$$

and this further implies that

$$\begin{aligned}
\log(4/\delta_0) &\le \log\left(4\left(\frac{(8\beta')^{\beta'} K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'}}{\Delta_S^*}\right)^{4/(3(1-\zeta))}\right) \\
&\le \log\left(\left(\frac{4^{1-\zeta}(8\beta')^{\beta'} K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'}}{\Delta_S^*}\right)^{4/(3(1-\zeta))}\right) \\
&= \frac{4}{3(1-\zeta)} \log\left(\frac{4^{1-\zeta}(8\beta')^{\beta'} K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'}}{\Delta_S^*}\right)
\end{aligned}$$

Plug this inequality back to the $\log(4/\delta_0)$ term in (26), and we have

$$\frac{1}{\delta_0^{1-\zeta}} = \frac{K_{\alpha,\beta,\zeta}\log(4/\delta_0)^{\beta'}}{\Delta_S^*} \le \frac{K_{\alpha,\beta,\zeta}}{\Delta_S^*} \cdot \left(\frac{4}{3(1-\zeta)}\right)^{\beta'} \log\left(\frac{4^{1-\zeta}(8\beta')^{\beta'} K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'}}{\Delta_S^*}\right)^{\beta'}$$

We write $D_{\alpha,\beta,\zeta} = 4^{1-\zeta}(8\beta')^{\beta'}K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'}$, $L_{\alpha,\beta,\zeta} = (4/(3(1-\zeta)))^{\beta'}K_{\alpha,\beta,\zeta}$, then we have

$$\delta_0^{1-\zeta} \geq \frac{\Delta_S^*}{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}. \tag{28}$$

By definition of i_0 that it is the largest integer i that satisfies $\gamma_i \geq \delta_0$, we also have

$$\gamma_{i_0}^{1-\zeta} \geq \frac{\Delta_S^*}{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}.$$

Recall that $\gamma_i = i^{-(1/\zeta-1)(1-\alpha)}$, thus

$$i_0 \leq \left(\frac{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^{\frac{\zeta}{(1-\zeta)^2(1-\alpha)}}. \tag{29}$$

We set

$$k_0 = \left\lfloor \left(\left(\frac{x/3}{2+\zeta+(2(1/\zeta-1)(1-\alpha))^{-1}} \right)^{\zeta} - i_0 - 1 \right)_+ \right\rfloor, \tag{30}$$

then for any round $k \leq k_0$, recall that $\gamma_i = i^{-(1/\zeta-1)(1-\alpha)}$, then $t_i' = \lceil \gamma_i^{-1/(1-\alpha)} \rceil \leq 1 + i^{1/\zeta-1}$ and $T_i' = \lceil \gamma_i^{-2} \rceil \leq 1 + i^{2(1/\zeta-1)(1-\alpha)}$, the cumulative satisficing regret in Step 1 and 2 can be upper bounded by linear in t_i' and T_i' as

$$\begin{aligned}
\sum_{i=i_0+1}^{i_0+k} R_i &\leq \sum_{i=i_0+1}^{i_0+k} (t_i' + T_i') \leq 2k + \sum_{i=i_0+1}^{i_0+k} (i^{1/\zeta-1} + i^{2(1/\zeta-1)(1-\alpha)}) \\
&< 2k + \int_{i_0+1}^{i_0+k+1} (y^{1/\zeta-1} + y^{2(1/\zeta-1)(1-\alpha)})\mathrm{d}y \\
&= 2k + \left(\zeta y^{1/\zeta} + \frac{1}{2(1/\zeta-1)(1-\alpha)+1} y^{2(1/\zeta-1)(1-\alpha)+1} \right) \Big|_{i_0+1}^{i_0+k+1} \\
&\overset{\text{(i)}}{\leq} 2k + \left(\left(\zeta + \frac{1}{2(1/\zeta-1)(1-\alpha)} \right) y^{1/\zeta} \right) \Big|_{i_0+1}^{i_0+k+1} \\
&\overset{\text{(ii)}}{\leq} 2k^{1/\zeta} + \left(\zeta + \frac{1}{2(1/\zeta-1)(1-\alpha)} \right) (i_0+k+1)^{1/\zeta} \\
&\leq \left(2 + \zeta + \frac{1}{2(1/\zeta-1)(1-\alpha)} \right) (i_0+k+1)^{1/\zeta} \\
&\overset{\text{(iii)}}{\leq} x/3,
\end{aligned} \tag{31}$$

where step (i) is because $2(1/\zeta-1)(1-\alpha)+1 \leq 1/\zeta$ for $1/2 \leq \alpha < 1$, step (ii) is because $y^{1/\zeta}\big|_{i_0+1} > 0$ and $k \leq k^{1/\zeta}$ for $\zeta < 1$ and $k \geq 1$, and step (iii) is based on $k < k_0$, $(i_0+k+1)^{1/\zeta}$ increasing in k, and the definition (30) of k_0. Therefore if the cumulative satisficing regret in Step 1 and 2 exceeds $x/3$, the total number of rounds should be at least $i_0 + k_0 + 1$. Furthermore, by Proposition 4 we know the probability that the total number of rounds exceeds $i_0 + k_0$ is at most 4^{-k_0}. Therefore,

plug into the definition (30) of k_0, and the probability of exceeding $x/3$ in Step 1 and 2 can be further bounded as

$$\begin{aligned}
&\Pr\left(\sum_{i=i_0+1}^{\infty} R_i > \frac{x}{3}\right) \le 4^{-k_0} \\
\le& 4^{i_0+2} \cdot 4^{\left(-\left(\frac{x/3}{2+\zeta+(2(1/\zeta-1)(1-\alpha))^{-1}}\right)^{\zeta}\right)} \\
\le& \left(e^2\right)^{i_0+2} \cdot \exp\left(-\left(\frac{x/3}{2+\zeta+(2(1/\zeta-1)(1-\alpha))^{-1}}\right)^{\zeta}\right) \\
\le& \exp\left(4+2i_0-\left(\frac{x}{9+(2(1-\zeta)/3)^{-1}(1-\alpha)^{-1}}\right)^{\zeta}\right),
\end{aligned} \tag{32}$$

where the final inequality is because $2+\zeta<3$ and $\zeta^{-1}-1>1-\zeta$.

Then we bound the probability that cumulative satisficing regret incurred in the Step 3 of all rounds exceeds $x/3$. For each round i, the probability that round i is not terminated after pulling non-satisficing arm $\hat{A}_i$, *i.e.*, $r(\hat{A}_i)<S$, for k_i times (in addition to the first T_i' times) occurs with probability at most

$$\begin{aligned}
&\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i'+k} - \sqrt{\frac{k_i^{\zeta}+\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{T_i'+k_i}} > S\right) \\
\le& \Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i'+k} - \sqrt{\frac{k_i^{\zeta}+\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{T_i'+k_i}} > r(\hat{A}_i)\right) \\
\le& \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k_i^{\zeta}\right),
\end{aligned} \tag{33}$$

where the second inequality is based on Hoeffding's inequality. The probability that satisficing regret incurred in the Step 3 in round i exceeds $x/3$ can be bounded by

$$\Pr\left(\max\left\{S-r(\hat{A}_i),0\right\}\cdot k_i > x/3\right) \le \Pr\left(k_i > x/3\right) \le \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-(x/3)^{\zeta}\right),$$

where the first inequality is because $\max\{S-r(\hat{A}_i),0\}\le 1$ based on reward $r\in[0,1]$ and $S\le r$, the second inequality is by applying (33). Note that $\zeta/\Gamma(\zeta^{-1})\le 3/2$, using Proposition 5, the probability

of total regret in LCB test exceeds $x/3$ can be bounded by

$$\begin{aligned}
&\Pr\left(\sum_{i=1}^{\infty}\max\left\{S-r(\hat{A}_i),0\right\}\cdot k_i > x/3\right)\\
\leq&\Pr\left(\sum_{i=1}^{i_0+N(1/4)}\max\left\{S-r(\hat{A}_i),0\right\}\cdot k_i > x/3\right)\\
\leq&\frac{2^{i_0+1}(1-1/4)}{1-2\cdot 1/4}\cdot\exp\left(-\frac{(x/3)^{\zeta}}{\zeta/(8\Gamma(\zeta^{-1}))+1}\right)\\
\leq&2^{i_0+2}\cdot\exp\left(-\frac{x^{\zeta}/3}{3/16+1}\right)\\
\leq&\exp\left(i_0+2-\frac{x^{\zeta}}{4}\right)
\end{aligned}\tag{34}$$

where the first inequality is based on Proposition 4, which says that for each round $i > i_0$, it ends and starts round $i+1$ with probability at most $1/4$. The second inequality is due to Proposition 5 where the total number of rounds follow a geometric distribution with success rate $3/4 = 1-1/4$ in addition to the initial i_0 rounds. The third inequality is because $\zeta < 1$ and $\Gamma(\zeta^{-1}) \geq 2/3$, thus $\zeta/8\Gamma(\zeta^{-1}) \leq 3/16$. The fourth inequality is due to $3(3/16+1) \leq 4$ and $2^{i_0+1} \leq e^{i_0+1}$.

As we assume $x > x_0 = 6i_0 + 6\zeta(i_0+1)^{1/\zeta}$, recall that $\gamma_i = i^{-(1/\zeta-1)(1-\alpha)}$, then $T_i' = \lceil \gamma_i^{-2} \rceil \leq t_i'$ and $t_i' = \lceil \gamma_i^{-1/(1-\alpha)} \rceil \leq 1 + i^{1/\zeta-1}$, then

$$\begin{aligned}
&\sum_{i=1}^{i_0}(t_i'+T_i') < \sum_{i=1}^{i_0}2t_i' < 2i_0 + \sum_{i=1}^{i_0}2i^{1/\zeta-1}\\
<&2i_0 + 2\int_1^{i_0+1} y^{1/\zeta-1}\mathrm{d}y = 2i_0 + 2\zeta y^{1/\zeta}\Big|_1^{i_0+1}\\
<&2i_0 + 2\zeta(i_0+1)^{1/\zeta} < x/3.
\end{aligned}$$

Therefore $x/3 > \sum_{i=1}^{i_0}(t_i'+T_i')$, then the satisficing regret incurred before round i_0 in Step 1 and 2 is always bounded by $x/3$. Recall i_0 is a constant without x or T, we can bound the tail of total regret with a combination of (31), (32), and (34) as

$$\begin{aligned}
\Pr\left(\texttt{Regret}_S > x\right) \leq&\Pr\left(\sum_{i=1}^{i_0}R_i > \frac{x}{3}\right) + \Pr\left(\sum_{i=i_0+1}^{\infty}R_i > \frac{x}{3}\right) + \Pr\left(\sum_{i=1}^{\infty}\max\left\{S-r(\hat{A}_i),0\right\}\cdot k_i > x/3\right)\\
\leq&\exp\left(4+2i_0-\left(\frac{x}{9+(2(1-\zeta)/3)^{-1}(1-\alpha)^{-1}}\right)^{\zeta}\right) + \exp\left(i_0+2-\frac{x^{\zeta}}{4}\right).
\end{aligned}$$

Finally, since $(x-x_0)_+ \leq x$, by replacing x with $(x-x_0)_+$ we obtain the probability bound in the theorem.

E.1. Proof of Proposition 4

By Condition 1 and similar to (18), because the expression of t_i' in γ_i remains unchanged compared to that in Algorithm 1, we still have

$$\mathbb{E}[r(A^*) - r(\hat{A}_i)] = \frac{1}{t_i'}\mathbb{E}\left[t_i' r(A^*) - \sum_{s=1}^{t_i'} r(A_s)\right] \leq \frac{1}{t_i'} C_1 t_i'^{\alpha} \log(t_i')^{\beta} \leq \frac{2C_1}{(1-\alpha)^{\beta}} \gamma_i \log(4/\gamma_i)^{\beta},$$

where the last inequality holds because $t_i' = \lceil 1/{\gamma_i}^{1/(1-\alpha)} \rceil \leq 2/{\gamma_i}^{1/(1-\alpha)} \leq 4/{\gamma_i}^{1/(1-\alpha)}$.

According to the condition of this proposition, we have i satisfies the condition

$$\frac{32C_1(1-\zeta)^{-\zeta/2}}{(1-\alpha)^{\beta}} \sqrt{2 + 2\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)} \gamma_i^{1-\zeta} \log(4/\gamma_i)^{\max\{\beta, 1/2\}} \leq \Delta_S^*.$$

Note that for any $0 < \zeta < 1$, $(1-\zeta)^{-\zeta/2} > 1$ and $\gamma_i^{1-\zeta} > \gamma_i$. Therefore i satisfies

$$\frac{32C_1}{(1-\alpha)^{\beta}} \gamma_i \log(4/\gamma_i)^{\max\{\beta, 1/2\}} \leq \Delta_S^*.$$

By the Markov's inequality and $\log(4/\gamma_i) \geq \log(4) > 1$ for $\gamma_i < 1$, we have

$$\Pr(r(A^*) - r(\hat{A}_i) \geq \Delta_S^*/2) \leq \Pr\left(r(A^*) - r(\hat{A}_i) \geq \frac{16C_1}{(1-\alpha)^{\beta}} \gamma_i \log(4/\gamma_i)^{\beta}\right) \leq \frac{1}{8}. \tag{35}$$

Now, if $r(\hat{A}_i) \geq r(A^*) - \Delta_S^*/2$, which happens with probability at least 7/8 according to the above inequality, then $r(\hat{A}_i) \geq S + \Delta_S^*/2$ holds. On the other hand, if i satisfies the condition stated in the proposition, then

$$(1-\zeta)^{-\zeta/2} \cdot \sqrt{2 + 2\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)} \gamma_i^{1-\zeta} \log(4/\gamma_i)^{1/2} \leq \Delta_S^*/16. \tag{36}$$

Note that for $f(y) = y^{\zeta}/(T_i' + y)$, it reaches its maximum when

$$\frac{\mathrm{d}}{\mathrm{d}y} f(y) = \frac{\zeta y^{\zeta-1} T_i' + y^{\zeta}(\zeta - 1)}{(T_i' + y)^2} = 0.$$

Therefore the maximum of $f(y)$ is reached at $\zeta T_i'/(1-\zeta)$. For simplicity, we set $D = \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)$. For $k = 0$, the confidence radius is bounded by

$$\sqrt{\frac{D}{T_i'}} \leq \sqrt{D}\gamma_i \leq 2\sqrt{D}\gamma_i^{1-\zeta} \log(4/\gamma_i)^{1/2} \leq \frac{\Delta_S^*}{8},$$

where the first inequality is because $T_i' = \lceil \gamma_i^{-2} \rceil \geq \gamma_i^{-2}$, the second inequality is because $\gamma_i < 1$ and $2\log(4/\gamma_i)^{1/2} > 1$, and the final inequality is due to (36). For $k \geq 1$, the confidence radius is bounded by

$$\begin{aligned}\sqrt{\frac{k^\zeta + D}{T_i' + k}} &\overset{\text{(i)}}{\leq} \sqrt{\frac{(D+1)k^\zeta}{T_i' + k}} \leq \sqrt{\frac{(D+1)\left(\frac{\zeta T_i'}{1-\zeta}\right)^\zeta}{T_i' + \left(\frac{\zeta T_i'}{1-\zeta}\right)}} \\ &\leq \sqrt{\frac{(D+1)(1-\zeta)^{-\zeta} \cdot \zeta^\zeta T_i'^\zeta}{T_i'}} \\ &\overset{\text{(ii)}}{\leq} \sqrt{(D+1)}(1-\zeta)^{-\zeta/2}\gamma_i^{1-\zeta} \\ &\overset{\text{(iii)}}{\leq} 2\sqrt{(D+1)}(1-\zeta)^{-\zeta/2}\gamma_i^{1-\zeta}\log(4/\gamma_i)^{1/2} \leq \frac{\Delta_S^*}{8}.\end{aligned} \tag{37}$$

where inequality (i) is because $k^\zeta \geq 1$ as k is a positive integer, inequality (ii) is because $\zeta^\zeta < 1$ as $\zeta < 1$ and $T_i' = \lceil \gamma_i^{-2} \rceil \geq \gamma_i^{-2}$ as $\gamma_i^{-1} = i^{(1/\zeta - 1)(1-\alpha)} \geq 1$, inequality (iii) is because $2\log(4/\gamma_i)^{1/2} > 1$ for any $\gamma_i < 1$, and the final inequality is due to (36). Suppose round i is terminated for some k, then

$$\frac{\hat{r}_i^{\text{tot}}}{T_i' + k} - \sqrt{\frac{k^\zeta + D}{T_i' + k}} < S = r(A^*) - \Delta_S^* \leq r(\hat{A}_i) - \frac{\Delta_S^*}{2} \leq r(\hat{A}_i) - 2\sqrt{\frac{k^\zeta + D}{T_i' + k}},$$

or equivalently

$$\frac{\hat{r}_i^{\text{tot}}}{T_i' + k} \leq r(\hat{A}_i) - \sqrt{\frac{k^\zeta + D}{T_i' + k}}.$$

By the Hoeffding's inequality we have

$$\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i' + k} \leq r(\hat{A}_i) - \sqrt{\frac{k^\zeta + D}{T_i' + k}}\right) \leq \exp(-k^\zeta - D) = \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^\zeta\right).$$

Therefore conditioned on $r(\hat{A}_i) \geq r(A^*) - \Delta_S^*/2$ the probability that round i is terminated by the end of the horizon is bounded by

$$\begin{aligned}&\sum_{k=1}^{\infty} \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^\zeta\right) \leq \frac{\zeta}{8\Gamma(\zeta^{-1})}\int_0^\infty \exp\left(-x^\zeta\right)\mathrm{d}x \\ \leq &\frac{\zeta}{8\Gamma(\zeta^{-1})} \cdot \zeta^{-1}\Gamma(\zeta^{-1}) = \frac{1}{8}.\end{aligned} \tag{38}$$

Combining (35) and (38) we conclude that the probability that round i terminates by the end of the time horizon is upper bounded by $1/4$.

E.2. Proof of Proposition 5

We define $Y_i = Q_i^\zeta$ for all i, then for all $i \in \mathbb{Z}^+$,

$$\Pr(Y_i \geq x) = \Pr(Q_i \geq x^{1/\zeta}) \leq K_1 \exp(-x).$$

For all $i \in \mathbb{Z}^+$, we have

$$\Pr\left(\exp\left(\frac{Y_i}{K_1+1}\right) \geq x\right) = \Pr(Y_i \geq (K_1+1)\log(x)) \leq K_1 x^{-K_1-1}.$$

Since $Q_i \geq 0$, we have $\exp(Y_i/(K_1+1)) \geq 1$, therefore

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{Y_i}{K_1+1}\right)\right] &= 1 + \int_1^\infty \Pr\left(\exp\left(\frac{Y_i}{K_1+1}\right) \geq x\right) \mathrm{d}x \\ &\leq 1 + \int_1^\infty K_1 x^{-K_1-1} \mathrm{d}x = 2.\end{aligned} \tag{39}$$

Let $N(\lambda)$ be a geometric random variable with success probability $1-\lambda$. Then we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\sum_{i=1}^{N(\lambda)} \frac{Y_i}{K_1+1}\right)\right] &\leq \sum_{n=1}^\infty \Pr(N(\lambda)=n) \cdot \mathbb{E}\left[\exp\left(\sum_{i=1}^{n} \frac{Y_i}{K_1+1}\right)\right] \\ &\leq \sum_{n=1}^\infty \Pr(N(\lambda)=n) \cdot \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{Y_i}{K_1+1}\right)\right] \\ &\leq \sum_{n=1}^\infty \Pr(N(\lambda)=n) \cdot 2^n \\ &= \sum_{n=1}^\infty \lambda^{n-1}(1-\lambda) \cdot 2^n = \frac{1-\lambda}{\lambda} \sum_{n=1}^\infty (2\lambda)^n \\ &= \frac{2-2\lambda}{1-2\lambda}.\end{aligned}$$

where the first inequality is by conditioning on all $n \geq 1$, the second inequality is by the independence of Y_i, the third inequality is by (39), and the final equality is based on the assumption that $\lambda < 1/2$. We define $W = \sum_{i=1}^{N+N(\lambda)} Y_i$, then combining the above inequality with (39) and the independence of y_i, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{W}{K_1+1}\right)\right] =& \mathbb{E}\left[\exp\left(\sum_{i=1}^{N+N(\lambda)} \frac{Y_i}{K_1+1}\right)\right] \\ =& \left(\mathbb{E}\left[\exp\left(\frac{Y_1}{K_1+1}\right)\right]\right)^N \cdot \mathbb{E}\left[\exp\left(\sum_{i=1}^{N(\lambda)} \frac{Y_i}{K_1+1}\right)\right] \\ \leq& \frac{2^{N+1}(1-\lambda)}{1-2\lambda}.\end{aligned}$$

By Markov's inequality and the above inequality on $\mathbb{E}[\exp(W/(K_1+1))]$, we have

$$\begin{aligned}\Pr(W \geq x) &= \Pr\left(\exp\left(\frac{W}{K_1+1}\right) \geq \exp\left(\frac{x}{K_1+1}\right)\right) \\ &\leq \mathbb{E}\left[\exp\left(\frac{W}{K_1+1}\right)\right] / \exp\left(\frac{x}{K_1+1}\right) \\ &= \frac{2^{N+1}(1-\lambda)}{1-2\lambda} \cdot \exp\left(-\frac{x}{K_1+1}\right).\end{aligned} \tag{40}$$

Since $0 < \zeta < 1$, we have

$$W = \sum_{i=1}^{N+N(\lambda)} Y_i = \sum_{i=1}^{N+N(\lambda)} Q_i^{\zeta} \geq \left(\sum_{i=1}^{N+N(\lambda)} Q_i \right)^{\zeta}.$$

Therefore, plugging into (40), we have

$$\begin{aligned} \Pr\left(\sum_{i=1}^{N+N(\lambda)} Q_i \geq x \right) &= \Pr\left(\left(\sum_{i=1}^{N+N(\lambda)} Q_i \right)^{\zeta} \geq x^{\zeta} \right) \\ &\leq \Pr(W \geq x^{\zeta}) \\ &\leq \frac{2^{N+1}(1-\lambda)}{1-2\lambda} \cdot \exp\left(-\frac{x^{\zeta}}{K_1 + 1} \right). \end{aligned}$$

F. Proof of Theorem 7

We first use the following proposition to bound the expected satisficing regret of each round.

Proposition 6. *If $r(A^*) \geq S$, then the satisficing regret incurred in round i is bounded by*

$$\frac{8C_1}{(1-\alpha)^{\beta}} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^{\beta}.$$

Proof of Proposition 6 is provided in Section F.1. We also recall the bound on i_0 is defined as the largest i such that $K_{\alpha,\beta,\zeta} \gamma_i^{1-\zeta} \log(4/\gamma_i)^{\max\{\beta,1/2\}} > \Delta_S^*$ where $K_{\alpha,\beta,\zeta} := \frac{32C_1(1-\zeta)^{-\zeta/2}}{(1-\alpha)^{\beta}} \sqrt{2 + 2\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}$. i_0 is upper bounded in (29) that

$$i_0 \leq \left(\frac{L_{\alpha,\beta,\zeta} \log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*} \right)^{\frac{\zeta}{(1-\zeta)^2(1-\alpha)}},$$

where $\beta' = \max\{\beta, 1/2\}$, $D_{\alpha,\beta,\zeta} = 4^{1/(1-\zeta)}(8\beta')^{\beta'} K_{\alpha,\beta,\zeta}/(1-\zeta)$, and $L_{\alpha,\beta,\zeta} = (4(1-\zeta)/3)^{\beta'} K_{\alpha,\beta,\zeta}$. Furthermore, recall that $\delta_0 \in (0,1)$ is the smallest solution to Equation (26) as

$$\delta_0^{1-\zeta} \log(4/\delta_0)^{\beta'} = \frac{\Delta_S^*}{K_{\alpha,\beta,\zeta}},$$

then we know $\delta_0 \leq \gamma_{i_0} = i_0^{-(1/\zeta - 1)(1-\alpha)}$.

Using the result of Proposition 6, we have that the satisficing regret by round i_0 is bounded by

$$\begin{aligned} &\sum_{i=1}^{i_0} \frac{8C_1}{(1-\alpha)^{\beta}} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^{\beta} \\ \leq &\sum_{i=1}^{i_0} \frac{8C_1}{(1-\alpha)^{\beta}} i^{\alpha(1-\zeta)/\zeta} \log(4/\delta_0)^{\beta} \\ \leq &\frac{8C_1}{(1-\alpha)^{\beta}} i_0^{1+\alpha(1-\zeta)/\zeta} \log(4 i_0^{(1/\zeta-1)(1-\alpha)})^{\beta} \\ \leq &\frac{8C_1}{(1-\alpha)^{\beta}(1-\zeta)^{\beta}} \left(\frac{L_{\alpha,\beta,\zeta} \log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*} \right)^{\frac{\zeta+\alpha(1-\zeta)}{(1-\zeta)^2(1-\alpha)}} \log\left(\frac{4L_{\alpha,\beta,\zeta} \log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*} \right)^{\beta}, \end{aligned} \tag{41}$$

where the first inequality is based on the definition $\gamma_i = i^{-(1/\zeta-1)(1-\alpha)}$ and that $\gamma_i = 2^{-i(1-\alpha)/\alpha} \geq \gamma_{i_0} \geq \delta_0$ for all $i \leq i_0$, the second inequality is because $i^{\alpha(1-\zeta)/\zeta} \leq i_0^{\alpha(1-\zeta)/\zeta}$ for any $i \in [i_0]$, and the third inequality is simply plugging into the upper bound on i_0 in (29).

By the results of Proposition 4, for every $k \geq 1$, the probability that round $i = i_0 + k$ starts within the time horizon is at most 4^{-k+1}. Therefore the regret incurred after round i_0 is bounded by

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{8C_1}{(1-\alpha)^\beta}(i_0+k)^{\alpha(1-\zeta)/\zeta}\log(4/\gamma_i)^\beta \cdot 4^{-k+1} \\
\leq &\frac{8C_1}{(1-\alpha)^\beta}\left(\sum_{k=1}^{i_0}(2i_0)^{\alpha(1-\zeta)/\zeta}\log(4/\gamma_i)^\beta \cdot 4^{-k+1} + \sum_{k=i_0+1}^{\infty}(2k)^{\alpha(1-\zeta)/\zeta}\log(4/\gamma_i)^\beta \cdot 4^{-k+1}\right) \\
\leq &\frac{8C_1}{(1-\alpha)^\beta}\left(i_0 \cdot (2i_0)^{\alpha(1-\zeta)/\zeta}\log(4i_0^{(1/\zeta-1)(1-\alpha)})^\beta + \sum_{k=1}^{\infty}(2k)^{\alpha(1-\zeta)/\zeta}\log(4k^{(1/\zeta-1)(1-\alpha)})^\beta \cdot 4^{-k+1}\right) \\
\leq &\frac{8C_1 \cdot 2^{\alpha(1-\zeta)/\zeta}}{(1-\alpha)^\beta(1-\zeta)^\beta}\left(\frac{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^{\frac{\zeta+\alpha(1-\zeta)}{(1-\zeta)^2(1-\alpha)}}\log\left(\frac{4L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^\beta + M,
\end{aligned} \tag{42}$$

where $M = \frac{8C_1}{(1-\alpha)^\beta}\sum_{k=1}^{\infty}(2k)^{\alpha(1-\zeta)/\zeta}\log(4k^{\alpha(1-\zeta)\zeta})^\beta \cdot 4^{-k+1}$ is a constant only dependent on α, β, ζ because 4^{-k+1} term decreases exponentially although $k^{\alpha(1-\zeta)/\zeta}$ grows in polynomial, and the final inequality is again simply plugging into the upper bound on i_0. Then, combining the satisficing regret before round i_0 (41) and after round i_0 (42), the total satisficing regret is bounded by

$$\begin{aligned}
\mathbb{E}[\texttt{Regret}_S] \leq M + &\frac{C_1}{(1-\alpha)^\beta(1-\zeta)^\beta}\left(2^{\alpha(1-\zeta)/\zeta}+2\right)\left(\frac{L_{\alpha,\beta,\zeta}^{\beta'}}{\Delta_S^*}\right)^{\frac{\zeta+\alpha(1-\zeta)}{(1-\zeta)^2(1-\alpha)}} \\
&\cdot \log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\frac{\beta'(\zeta+\alpha(1-\zeta))}{(1-\zeta)^2(1-\alpha)}}\log\left(\frac{4L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^\beta.
\end{aligned} \tag{43}$$

Furthermore, since the length of the time horizon is T time steps, we have

$$T \geq \sum_{i=1}^{i_{\max}-1} t_i' \geq \sum_{i=1}^{i_{\max}-1} i^{1/\zeta-1} \geq \int_0^{i_{\max}-1} x^{1/\zeta-1}\mathrm{d}x = \zeta(i_{\max}-1)^{1/\zeta},$$

thus the maximum number of rounds is bounded by $i_{\max} \leq 1 + \lfloor (T/\zeta)^\zeta \rfloor = \lceil (T/\zeta)^\zeta \rceil$ because $i_{\max}$ is an integer. Recall that $\gamma_i = i^{-(1/\zeta-1)(1-\alpha)}$ is decreasing in i, for any $i \leq i_{\max}$, we have

$$\begin{aligned}
\gamma_i \geq &\gamma_{i_{\max}} = i_{\max}^{-(1/\zeta-1)(1-\alpha)} \\
\geq &(2^{1/\zeta}T/\zeta)^{-(1-\zeta)(1-\alpha)} \\
\geq &(2^{1/\zeta}T/\zeta)^{-1},
\end{aligned} \tag{44}$$

where the the second inequality is because $i_{\max} \le \lceil (T/\zeta)^\zeta \rceil \le (T/\zeta)^\zeta + 1 \le 2(T/\zeta)^\zeta$ based on $T/\zeta \ge 1$ for $T \ge 1$ and $\zeta < 1$, and third inequality is because $\zeta < 1$ and $1/2 \le \alpha < 1$. So, $4/\gamma_i \le 2^{2+1/\zeta} T/\zeta$. By the results of Proposition 6, we have that the satisficing regret is bounded by

$$\begin{aligned}
\mathbb{E}[\texttt{Regret}_S] \le & \sum_{i=1}^{i_{\max}} \frac{8C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^\beta \\
\overset{\text{(i)}}{\le} & \sum_{i=1}^{i_{\max}} \frac{8C_1}{(1-\alpha)^\beta} i^{\alpha(1-\zeta)/\zeta} \log(2^{2+1/\zeta} T/\zeta)^\beta \\
\le & \frac{8C_1}{(1-\alpha)^\beta} \log(2^{2+1/\zeta} T/\zeta)^\beta \int_1^{i_{\max}+1} y^{\alpha(1-\zeta)/\zeta} \mathrm{d}y \\
= & \frac{8C_1}{(1-\alpha)^\beta (\alpha(1-\zeta)/\zeta + 1)} y^{\alpha(1-\zeta)/\zeta+1} \Big|_1^{i_{\max}+1} \log(2^{2+1/\zeta} T/\zeta)^\beta \\
\le & \frac{8C_1}{(1-\alpha)^\beta (\alpha(1-\zeta)/\zeta + 1)} (i_{\max}+1)^{\alpha(1-\zeta)/\zeta+1} \log(2^{2+1/\zeta} T/\zeta)^\beta \\
\overset{\text{(ii)}}{\le} & \frac{8C_1 \zeta^{-\alpha(1-\zeta)-\zeta} \cdot 4^{\alpha(\zeta^{-1}-1)+1}}{(1-\alpha)^\beta (\alpha(\zeta^{-1}-1)+1)} T^{\alpha(1-\zeta)+\zeta} \log(2^{2+1/\zeta} T/\zeta)^\beta.
\end{aligned} \tag{45}$$

where step (i) is based on (44) and step (ii) is because $i_{\max} + 1 \le 2 i_{\max} \le 2(T/\zeta)^\zeta + 2 \le 4(T/\zeta)^\zeta$ because $i_{\max} \ge 1$ and $(T/\zeta)^\zeta \ge 1$.

Combining the two parts of results, *i.e.*, the constant satisficing regret (43) and the sub-linear satisficing regret (45), we have completed the proof of Theorem 7 by putting all constant terms into the polylog term.

F.1. Proof of Proposition 6

Following the proof pattern of Proposition 1, as the expression of t_i' in γ_i remains $t_i' = \lceil \gamma_i^{-1/(1-\alpha)} \rceil$, the same as (14), the satisficing regret incurred when running $\texttt{ALG}(t_i')$, *i.e.*, satisficing regret incurred in Step 1, is still bounded with Condition 1 by

$$\mathbb{E}\left[\sum_{s=1}^{t_i'} \max\{S - r(A_s), 0\}\right] \le \mathbb{E}\left[t_i' r(A^*) - \sum_{s=1}^{t_i'} r(A_s)\right] \le \frac{2C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^\beta.$$

Since $\hat{A}_i$ is selected randomly according to the trajectory of $\texttt{ALG}(t_i')$ in round i, the same as (15), we still have

$$\mathbb{E}[\max\{S - r(\hat{A}_i), 0\}] \le \mathbb{E}[r(A^*) - r(\hat{A}_i)] \le \frac{2C_1}{(1-\alpha)^\beta} \gamma_i \log(4/\gamma_i)^\beta.$$

Therefore the satisficing regret incurred when pulling $\hat{A}_i$ for $T_i' = \lceil \gamma_i^{-2} \rceil$ times in Step 2, similar to (16), is still bounded by

$$\left\lceil \frac{1}{\gamma_i^2} \right\rceil \cdot \frac{2C_1}{(1-\alpha)^\beta} \gamma_i \log(4/\gamma_i)^\beta \le \frac{4C_1}{(1-\alpha)^\beta} \gamma_i^{-1} \log(4/\gamma_i)^\beta \le \frac{4C_1}{(1-\alpha)^\beta} \gamma_i^{-\alpha/(1-\alpha)} \log(4/\gamma_i)^\beta.$$

Recall (33), if $r(\hat{A}_i) < S$, then for every k, round i is not terminated after pulling arm $\hat{A}_i$ for k times (in addition to the first T_i' times) occurs with probability at most

$$\begin{aligned}
&\Pr\left(\frac{\hat{r}_i^{\mathrm{tot}}}{T_i'+k} - \sqrt{\frac{k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{T_i'+k}} > S\right) \\
\leq &\Pr\left(\frac{\hat{r}_i^{\mathrm{tot}}}{T_i'+k} - \sqrt{\frac{k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{T_i'+k}} > r(\hat{A}_i)\right) \\
\leq &\exp\left(-k^\zeta + \log\left(\frac{\zeta}{8\Gamma(\zeta^{-1})}\right)\right) = \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^\zeta\right).
\end{aligned}$$

Therefore, the same as to (38), expected number of arm pulls of $\hat{A}_i$ after first pulling it T_i' times, *i.e.*, the expected number of pulls in Step 3, is bounded by

$$\begin{aligned}
&\mathbb{E}\left[\text{Number of times } \hat{A}_i \text{ is pulled in round } i \text{ after the first } T_i' \text{ times}\right] \\
= &\sum_{s=1}^{\infty} \Pr(k \geq s) = \frac{\zeta}{8\Gamma(\zeta^{-1})}\sum_{s=1}^{\infty}\exp\left(-s^\zeta\right) \\
\leq &\frac{\zeta}{8\Gamma(\zeta^{-1})}\int_0^{\infty}\exp\left(-x^\zeta\right)\mathrm{d}x \\
\leq &\frac{\zeta}{8\Gamma(\zeta^{-1})}\cdot\zeta^{-1}\Gamma(\zeta^{-1}) = \frac{1}{8} \leq 1.
\end{aligned}$$

Summing up the satisficing regret from Step 1, Step 2, and Step 3, we have that the satisficing regret of round i is still bounded by

$$\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

Note that, this is exactly the same as (17) in Proposition 1

G. Proof of Theorem 8

We first prove that the regret incurred in round i is at most

$$\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

For Step 1, the regret incurred when running $\texttt{ALG}(t_i')$ is bounded by

$$\mathbb{E}\left[t_i' r(A^*) - \sum_{s=1}^{t_i'} r(A_s)\right] \leq C_1 t_i'^{\alpha}\log(t_i')^\beta \leq \frac{2C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta. \tag{46}$$

The same as (18), we further have

$$\mathbb{E}[r(A^*) - r(\hat{A}_i)] \leq \frac{2C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^\beta. \tag{47}$$

Therefore, for Step 2, the same as (16), the satisficing regret incurred when pulling $\hat{A}_i$ is bounded by

$$\left\lceil \frac{1}{\gamma_i^2} \right\rceil \cdot \frac{2C_1}{(1-\alpha)^\beta}\gamma_i \log(4/\gamma_i)^\beta \leq \frac{4C_1}{(1-\alpha)^\beta}\gamma_i^{-1}\log(4/\gamma_i)^\beta \leq \frac{4C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta. \quad (48)$$

Since $r(\hat{A}_i) \leq r(A^*) < S$, for every k, round i is not terminated with probability at most

$$\begin{aligned}
&\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i'+k} - \sqrt{\frac{k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{T_i'+k}} > S\right) \\
\leq &\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i'+k} - \sqrt{\frac{k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{T_i'+k}} > r(\hat{A}_i)\right) \\
\leq &\exp\left(-k^\zeta + \log\left(\frac{\zeta}{8\Gamma(\zeta^{-1})}\right)\right) = \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^\zeta\right).
\end{aligned}$$

Therefore, for Step 3, following the same calculation process in (38), expected number of arm pulls of $\hat{A}_i$ after pulling T_i' times is bounded by

$$\begin{aligned}
&\frac{\zeta}{8\Gamma(\zeta^{-1})}\sum_{s=1}^{\infty}\exp\left(-k^\zeta\right) \leq \frac{\zeta}{8\Gamma(\zeta^{-1})}\int_0^\infty \exp\left(-x^\zeta\right)\mathrm{d}x \\
\leq &\frac{\zeta}{8\Gamma(\zeta^{-1})}\cdot\zeta^{-1}\Gamma(\zeta^{-1}) = \frac{1}{8} \leq 1.
\end{aligned}$$

Then, the expected standard regret incurred in Step 3 is bounded by

$$1 \cdot \mathbb{E}[r(A^*) - r(\hat{A}_i)] \leq \frac{2C_1}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^\beta \leq \frac{2C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta, \quad (49)$$

where the first inequality is based on (47) and the second inequality is based on $\gamma_i < 1$.

Summing up the standard regret in Step 1 (46), 2(48), and 3(49), we have that the satisficing regret of round i is bounded by

$$\frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

Since the length of the time horizon is T time steps, we have

$$T \geq \sum_i^{i_{\max}} t_i' \geq \sum_i^{i_{\max}} i^{1/\zeta - 1} \geq \int_0^{i_{\max}} x^{1/\zeta-1}\mathrm{d}x = \zeta i_{\max}^{1/\zeta},$$

thus the maximum number of rounds is bounded by $i_{\max} \leq \lfloor (T/\zeta)^\zeta \rfloor$, which implies that $i_{\max}+1 \leq 2i_{\max} \leq (T/\zeta)^\zeta$ because $i_{\max} \geq 1$. Recall from (44) that $4/\gamma_i \leq 2^{2+1/\zeta}T/\zeta$ and follow the same chain of inequalities in (45), we have that the regret is bounded by

$$\begin{aligned}
\mathbb{E}[\texttt{Regret}] &\leq \sum_{i=1}^{i_{\max}} \frac{8C_1}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta \\
&\leq \sum_{i=1}^{i_{\max}} \frac{8C_1}{(1-\alpha)^\beta} i^{\alpha(1-\zeta)/\zeta}\log(2^{2+1/\zeta}T/\zeta)^\beta
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{8C_1}{(1-\alpha)^\beta} \log(2^{2+1/\zeta}T/\zeta)^\beta \int_1^{i_{\max}+1} y^{\alpha(1-\zeta)/\zeta} \mathrm{d}y \\
&= \frac{8C_1}{(1-\alpha)^\beta(\alpha(1-\zeta)/\zeta+1)} y^{\alpha(1-\zeta)/\zeta+1}\Big|_1^{i_{\max}+1} \log(2^{2+1/\zeta}T/\zeta)^\beta \\
&\leq \frac{8C_1}{(1-\alpha)^\beta(\alpha(1-\zeta)/\zeta+1)} (i_{\max}+1)^{\alpha(1-\zeta)/\zeta+1} \log(2^{2+1/\zeta}T/\zeta)^\beta \\
&\leq \frac{8C_1\zeta^{-\alpha(1-\zeta)-\zeta}\cdot 4^{\alpha(\zeta^{-1}-1)+1}}{(1-\alpha)^\beta(\alpha(\zeta^{-1}-1)+1)} T^{\alpha(1-\zeta)+\zeta} \log(2^{2+1/\zeta}T/\zeta)^\beta
\end{aligned}
$$

Therefore the regret is bounded by $C_1 T^{\alpha(1-\zeta)+\zeta} \cdot \mathrm{polylog}(T)$.

H. Tail Risk Control for Satisficing in Finite-Armed Bandits

In Section 6, we established `SELECT-LITE`, which is able to attain a constant expected satisficing regret and a light-tailed satisficing regret distribution in the realizable case. However, in the non-realizable case, the standard regret distribution of `SELECT-LITE` may not be light-tailed. In this section, we focus on the finite-armed bandit setting and establish an algorithm that: 1) attains a constant expected satisficing regret and a light-tailed satisficing regret distribution in the realizable case; 2) attains a sub-linear expected standard regret and a light-tailed standard regret distribution in the non-realizable case.

Similar to Condition 1, we first assume access to a learning oracle that attains a sub-linear expected standard regret and a light-tailed standard regret distribution for finite-armed bandits.

Condition 2. *There exists an finite-armed bandit algorithm* `ALG`(t)*, functions* $\Lambda(K,t), \Lambda'(K,t)$ *and* $\zeta > 0$ *such that for any instance of* K*-arm bandits and any time horizon* t*,*

1. *The expected standard regret of* `ALG`(t) *is upper bounded by* $CK^{\alpha_1}t^{\alpha_2}\log(t)^\beta$*, where* $C, \alpha_1, \alpha_2, \beta$ *are constants independent from* K *and* t*;*
2. *The probability that the standard regret of* `ALG`(t) *exceeds* x *is upper bounded by* $\Lambda(K,t)\exp(-x^\zeta/\Lambda'(K,t))$*;*
3. $\Lambda(K,t)$ *can be at most polynomially dependent on* t *and* $\Lambda'(K,t)$ *can be at most poly-logarithmically dependent on* t*.*

Without loss of generality we also assume that $\Lambda(K,t)$ and $\Lambda'(K,t)$ are monotonically increasing in t for every K, and is monotonically decreasing in x for every fixed t and K. Condition 2 can be satisfied by existing algorithms, such as the one established in Simchi-Levi et al. (2025).

For finite-armed bandits, we notice that the oracle algorithm is able to output the candidate arm that asymptotically converges to the optimal arm by choosing the one that is the most frequently pulled. Because the arm is already been pulled by at least t_i/K times, leveraging this history already yields a tightened confidence interval for it, making forced sampling unnecessary. This holds for both realizable and non-realizable cases.

Therefore, we make three modifications to SELECT to make both satisficing regret distribution and standard regret distribution light-tailed for finite-armed bandits with the help of a light-tailed oracle algorithm. First, we get rid of forced sampling and directly start LCB test after running the oracle. Second, we use the arm that is pulled most frequently during $\texttt{ALG}(t_i)$ as the candidate arm, and the historical data of this arm being pulled when running $\texttt{ALG}(t_i)$ can be used in LCB test. Third, we use larger confidence intervals in the LCB test the same as the one in SELECT-LITE. The complete algorithm is provided in Algorithm 3.

Algorithm 3 Light-Tailed Satisficing Regret Minimization for Finite-Armed Bandits (SELECT-LITE+)

Input: time horizon T, satisficing level S, tail distribution parameter ζ
Set $\gamma_i \leftarrow 2^{-i(1-\alpha_2)/\alpha_2}$ for all $i = 1, 2, 3 \dots$
for round $i = 1, 2, \dots$ **do**
  Set $t_i \leftarrow \lceil \gamma_i^{-1/(1-\alpha_2)} \rceil$
  Run $\texttt{ALG}(t_i)$, let $\hat{A}_i$ denote the arm that is pulled most frequently along the trajectory during $\texttt{ALG}(t_i)$.
  Set τ_i be the number of times $\hat{A}_i$ is pulled during $\texttt{ALG}(t_i)$, $k \leftarrow 0$.
  Set $\hat{r}_i^{\text{tot}}$ be the sum of observed reward of pulling $\hat{A}_i$ for τ_i times during $\texttt{ALG}(t_i)$.
  while $LCB(\hat{A}_i) \geq S$ **do**
    Set $k \leftarrow k + 1$
    Pull arm $\hat{A}_i$ and observe Y_t, where t is the current time step
    Update $\hat{r}_i^{\text{tot}} \leftarrow \hat{r}_i^{\text{tot}} + Y_t$
    Set

$$LCB(\hat{A}_i) \leftarrow \frac{\hat{r}_i^{\text{tot}}}{\tau_i + k} - \sqrt{\frac{2k^{\zeta} + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{\tau_i + k}}$$

  end while
end for

With the help of a finite-armed bandit learning algorithm with light-tailed standard regret, we use the following theorem to show that SELECT-LITE+ can enjoy a light-tailed satisficing regret distribution in the realizable case and a light-tailed standard regret distribution in the non-realizable case.

Theorem 9. *If $r(A^*) \geq S$, then under Condition 2, for any $x > 0$ the probability that satisficing regret of SELECT-LITE+ exceeds x is bounded by*

$$\begin{aligned} &Pr(\texttt{Regret}_S > x) \\ \leq &(\alpha_2 \log_2(T) + 1)\Lambda(K,T) \exp\left(\frac{-x^{\zeta}}{2^{\zeta}(\alpha_2 \log_2(T) + 1)^{\zeta}\Lambda'(K,T)}\right) + \exp\left(i_0 + 1 - \frac{x^{\zeta}}{3}\right), \end{aligned}$$

where $i_0 = O\left(\log_2\left(1/\Delta_S^* \cdot \log(1/\Delta_S^*)^{\beta'}\right)\right)$ is a constant independent of x.

If $r(A^*) < S$ and the oracle algorithm `ALG` *is light-tailed, then the probability that standard regret of* `SELECT-LITE+` *exceeds x is bounded by*

$$\begin{aligned} &Pr(\texttt{Regret} > x) \\ \leq &(\alpha_2 \log_2(T) + 1)\Lambda(K,T) \exp\left(\frac{-x^\zeta}{2(\alpha_2 \log_2(T)+1)\Lambda'(K,T)}\right) + 2T^{\alpha_2} \exp\left(-(x/2)^\zeta\right). \end{aligned}$$

Proof of Theorem 9 is provided in Section H.1. In the next theorem, we show that similar to `SELECT`, the satisficing regret of `SELECT-LITE+` can be bounded by a constant in the realizable case.

Theorem 10. *If $r(A^*) \geq S$, then the satisficing regret of* `SELECT-LITE+` *is bounded by*

$$\mathbb{E}[\texttt{Regret}_S] \leq \min\left\{C_1 \left(\frac{C_1 K^{\alpha_1+1}}{\Delta_S^*}\right)^{\frac{\alpha_2}{(1-\zeta)(1-\alpha_2)}} \cdot \text{polylog}\left(\frac{C_1 K^{\alpha_1+1}}{\Delta_S^*}\right), C_1 T^{\alpha_2} \cdot \text{polylog}(T)\right\}.$$

Proof of Theorem 10 is provided in Section H.2. The proof relies on two critical results. In Proposition 7, we show an upper bound on the satisficing regret of round i. The proof is provided in Section H.3 of the appendix. In Proposition 8, we also show that once `SELECT-LITE+` runs for enough rounds, it is unlikely to start a new round. The proof is provided in Section H.4 of the appendix.

Proposition 7. *If $r(A^*) \geq S$, then the satisficing regret incurred in round i is upper bounded by*

$$\frac{4C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} \gamma_i^{-\alpha_2/(1-\alpha_2)} \log(4/\gamma_i)^\beta.$$

Proposition 8. *If $r(A^*) \geq S$, then round i terminates within the time horizon with probability at most $1/4$ if i that satisfies*

$$\frac{32 C_1 K^{\alpha_1+1}(1-\zeta)^{-\zeta/2}}{(1-\alpha_2)^\beta} \sqrt{2 + 4\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)} \gamma_i^{1-\zeta} \log(4/\gamma_i)^{\max\{\beta,1/2\}} \leq \Delta_S^*. \tag{50}$$

We also show that in the non-realizable case, `SELECT-LITE+` enjoys the same standard regret bound as the oracle shown in Theorem 11. The proof is provided in Section H.5 of the appendix.

Theorem 11. *If $r(A^*) < S$, the standard regret of* `SELECT-LITE+`*is bounded by*

$$C_1 K^{\alpha_1+1} T^{\alpha_2} \cdot \text{polylog}(T).$$

H.1. Proof of Theorem 9

For notational brevity we define $p(x;K,t) = \Lambda(K,t)\exp(-x^\zeta/\Lambda'(K,t))$. We first bound the tail probability of running the oracle. We define the satisficing regret incurred in the oracle algorithm in each round i as R_i. Since by definition of $t_i = \lceil 2^{i/\alpha_2} \rceil$, the number of rounds that end within

the time horizon is bounded by $2^{i/\alpha_2} \leq T$, *i.e.*, $i \leq \lfloor \alpha_2 \log_2(T) \rfloor$ because i is an integer. Then, the maximum number of rounds is bounded by $i_{\max} = \lceil \alpha_2 \log_2(T) \rceil$, and the probability that cumulative satisficing regret incurred in oracle algorithm exceeds $x/2$ is

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{i_{\max}} R_i > \frac{x}{2}\right) \leq & \sum_{i=1}^{i_{\max}} \Pr\left(R_i > \frac{x}{2i_{\max}}\right) \\
\leq & \sum_{i=1}^{i_{\max}} p\left(\frac{x}{2i_{\max}}; K, t_i\right) \\
\leq & i_{\max} p\left(\frac{x}{2i_{\max}}; K, T\right) \\
\leq & (\alpha_2 \log_2(T) + 1) p\left(\frac{x}{2(\alpha_2 \log_2(T) + 1)}; K, T\right)
\end{aligned} \tag{51}$$

The third inequality is because $p(x; K, t)$ is monotonically increasing in t, and the fourth inequality is because $p(x; K, t)$ is monotonically decreasing in x.

We then bound the tail probability of LCB test. This is divided into realizable and non-realizable cases. For the realizable case, the probability that satisficing regret incurred in the LCB test in round i exceeds $x/2$ can be bounded by

$$\begin{aligned}
& \Pr\left(\max\left\{S - r(\hat{A}_i), 0\right\} \cdot k_i > x/2\right) \\
\leq & \Pr\left(k_i > x/2\right) \\
\leq & \frac{\zeta}{8\Gamma(\zeta^{-1})} \exp\left(-(x/2)^{\zeta}\right),
\end{aligned}$$

where the first inequality is because $\max\{S - r(\hat{A}_i), 0\} \leq 1$ based on reward $r \in [0, 1]$ and $S \leq r$, and the second inequality is by applying (33). Note that $\zeta/\Gamma(\zeta^{-1}) \leq 3/2$, $\exp(2^{\zeta}) \leq e^2$. Proposition 8 tells that the total number of rounds follows a geometric distribution with success rate $3/4$ in addition to the initial i_0 rounds, where i_0 is the largest i that fails to satisfy Inequality (50) in Proposition 8. Then, using Proposition 5, similar to (34), the probability of total regret in the LCB test exceeds $x/2$ can be bounded by

$$\begin{aligned}
& \Pr\left(\sum_{i=1}^{\infty} \max\left\{S - r(\hat{A}_i), 0\right\} \cdot k_i > x/2\right) \\
\leq & \Pr\left(\sum_{i=1}^{i_0 + N(1/4)} \max\left\{S - r(\hat{A}_i), 0\right\} \cdot k_i > x/2\right) \\
\leq & \frac{2^{i_0}(1 - 1/4)}{1 - 2 \cdot 1/4} \cdot \exp\left(-\frac{(x/2)^{\zeta}}{\zeta/(8\Gamma(\zeta^{-1})) + 1}\right) \\
\leq & 2^{i_0 + 1} \cdot \exp\left(-\frac{x^{\zeta}/2}{3/16 + 1}\right) \\
\leq & \exp\left(i_0 + 1 - \frac{x^{\zeta}}{3}\right)
\end{aligned} \tag{52}$$

For the non-realizable case, the probability that cumulative satisficing regret incurred in oracle algorithm exceeds $x/2$ is upper bounded by

$$\begin{aligned}
&\Pr\left(\sum_{i=1}^{\infty}\max\left\{S-r(\hat{A}_i),0\right\}\cdot k_i > x/2\right)\\
\leq&\Pr\left(\left(\sum_{i=1}^{i_{\max}}k_i\right)^{\zeta} > \left(\frac{x}{2}\right)^{\zeta}\right)\\
=&\Pr\left(\exp\left(\sum_{i=1}^{i_{\max}}k_i^{\zeta}\right) > \exp\left(\left(\frac{x}{2}\right)^{\zeta}\right)\right)\\
\leq&\frac{\mathbb{E}\left[\exp\left(\sum_{i=1}^{i_{\max}}k_i^{\zeta}\right)\right]}{\exp((x/2)^{\zeta})}\\
=&\exp\left(-(x/2)^{\zeta}\right)\prod_{i=1}^{i_{\max}}\mathbb{E}\left[\exp\left(k_i^{\zeta}\right)\right],
\end{aligned} \tag{53}$$

where where the first inequality is because $\max\{S-r(\hat{A}_i),0\}\leq 1$ based on reward $r\in[0,1]$ and $S\leq r$, and the second inqualityi is based on Markov inequality, and the independence among each round i is used in the final equality. For each term $\mathbb{E}\left[\exp\left(k_i^{\zeta}\right)\right]$, it can be bounded by

$$\begin{aligned}
\mathbb{E}\left[\exp\left(k_i^{\zeta}\right)\right] &= \sum_{s=0}^{\infty}\Pr\left(k_i=s\right)\cdot\exp\left(s^{\zeta}\right)\\
&\leq 1+\sum_{s=1}^{\infty}\Pr\left(k_i=s\right)\cdot\exp\left(s^{\zeta}\right)\\
&\leq 1+\frac{\zeta}{8\Gamma(\zeta^{-1})}\sum_{s=1}^{\infty}\exp\left(-2s^{\zeta}+s^{\zeta}\right)\\
&\leq 1+\frac{\zeta}{8\Gamma(\zeta^{-1})}\cdot\zeta^{-1}\Gamma(\zeta^{-1})\leq 2,
\end{aligned} \tag{54}$$

where the second inequality is based on Inequality (61). Note that, the bound on $\mathbb{E}[\exp(k_i^{\zeta})]$ is independent of the round i. With $\max\{S-r(\hat{A}_i),0\}\leq 1$ based on reward $r\in[0,1]$ and $r(A^*)\geq S$, we plug (54) into (53), and can bound the probability that cumulative satisficing regret incurred in oracle algorithm exceeds $x/2$ by

$$\Pr\left(\sum_{i=1}^{\infty}\max\left\{S-r(\hat{A}_i),0\right\}\cdot k_i > x/2\right)\leq\exp\left(-(x/2)^{\zeta}\right)\cdot 2^{1+\alpha_2\log_2(T)} = 2T^{\alpha_2}\exp\left(-(x/2)^{\zeta}\right). \tag{55}$$

Combining oracle tail probability (51) with the tail probability (52) in the LCB test for the realizable case and the tail probability (55) in the LCB test for the non-realizable case, we bound the tail probability in the realizable case as

$$\Pr\left(\texttt{Regret}_S > x\right)\leq(\alpha_2\log_2(T)+1)p\left(\frac{x}{2(\alpha_2\log_2(T)+1)};K,T\right)+\exp\left(i_0+1-\frac{x^{\zeta}}{3}\right)$$

and tail probability in the non-realizable case as

$$\Pr\left(\texttt{Regret} > x\right) \leq (\alpha_2 \log_2(T)+1) p\left(\frac{x}{2(\alpha_2\log_2(T)+1)}; K, T\right) + 2T^{\alpha_2}\exp\left(-(x/2)^{\zeta}\right).$$

H.2. Proof of Theorem 10

Denote $\beta' = \max\{\beta, 1/2\}$ and $K_{\alpha,\beta,\zeta} = \frac{32C_1K^{\alpha_1+1}(1-\zeta)^{-\zeta/2}}{(1-\alpha_2)^{\beta}}\sqrt{2+4\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}$, and i_0 as

$$i_0 = \max\{i : K_{\alpha,\beta,\zeta}\gamma_i^{1-\zeta}\log(4/\gamma_i)^{\max\{\beta,1/2\}} > \Delta_S^*\}.$$

We first follow the proof of Theorem 6 in Section E and Theorem 1 in Section A to get an upper bound on i_0. Denote $\delta_0 \in (0,1)$ the smallest solution to the following equation

$$\delta_0^{1-\zeta}\log(4/\delta_0)^{\beta'} = \frac{\Delta_S^*}{K_{\alpha,\beta,\zeta}},$$

the same as Inequality (26). Following exactly the same proof steps to Inequality (27), we get

$$\frac{1}{\delta_0^{1-\zeta}} \leq \left(\frac{(8\beta')^{\beta'}K_{\alpha,\beta,\zeta}/(1-\zeta)^{\beta'}}{\Delta_S^*}\right)^{4/3}.$$

We write $D_{\alpha,\beta,\zeta} = 4^{1/(1-\zeta)}(8\beta')^{\beta'}K_{\alpha,\beta,\zeta}$, $L_{\alpha,\beta,\zeta} = (4(1-\zeta)/3)^{\beta'}K_{\alpha,\beta,\zeta}$, then, the same as (28), we have

$$\delta_0^{1-\zeta} \geq \frac{\Delta_S^*}{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}.$$

By definition of i_0 we also have

$$\gamma_{i_0}^{1-\zeta} \geq \frac{\Delta_S^*}{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}},$$

thus

$$i_0 \leq \frac{\alpha_2}{(1-\alpha_2)(1-\zeta)}\log_2\left(\frac{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right).$$

Using the result of Proposition 7 we have that the satisficing regret by round i_0 is bounded by

$$\begin{aligned}
&\sum_{i=1}^{i_0}\frac{4C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}}\gamma_i^{-\alpha_2/(1-\alpha_2)}\log(4/\gamma_i)^{\beta} \\
=&\sum_{i=1}^{i_0}\frac{4C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}}2^i\log(4/\delta_0)^{\beta} = \frac{4C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}}(2^{i_0}-1)\log(4/\delta_0)^{\beta} \qquad (56)\\
\leq&\frac{8C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}(1-\zeta)^{\beta}}\left(\frac{L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^{\frac{\alpha_2}{(1-\alpha_2)(1-\zeta)}}\log\left(\frac{2L_{\alpha,\beta,\zeta}\log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^{\beta}.
\end{aligned}$$

By the results of Proposition 8, for every $k \geq 1$, the probability that round $i_0 + k$ starts within the time horizon is at most 4^{-k+1}. Therefore the regret incurred after round i_0 is bounded by

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{4C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} 2^{i_0+k} \log(4/\gamma_{i_0+k})^\beta \cdot 4^{-k+1} \\
\leq &\sum_{k=1}^{\infty} \frac{16C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} 2^{i_0} \left(\log(4/\gamma_{i_0}) + \frac{1-\alpha_2}{\alpha_2} k\right)^\beta \cdot 2^{-k} \\
\leq &\sum_{k=1}^{\infty} \frac{16 \cdot 2^\beta C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} 2^{i_0} \log(4/\gamma_{i_0})^\beta \cdot \frac{(1-\alpha_2)^\beta}{\alpha_2^\beta} k^\beta \cdot 2^{-k} \\
\leq &\frac{16 \cdot 2^\beta C_1 K^{\alpha_1+1}}{\alpha_2^\beta} 2^{i_0} \log(4/\delta_0)^\beta \cdot \sum_{k=1}^{\infty} k^\beta \cdot 2^{-k} \\
\leq &\frac{16 \cdot 2^\beta C_1 K^{\alpha_1+1} M}{\alpha_2^\beta} 2^{i_0} \log\left(\left(4 \cdot \delta_0^{-1+\zeta}\right)^{1/(1-\zeta)}\right)^\beta \\
\leq &\frac{16 \cdot 2^\beta C_1 K^{\alpha_1+1} M}{\alpha_2^\beta (1-\zeta)^\beta} \left(\frac{L_{\alpha,\beta,\zeta} \log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^{\frac{\alpha_2}{(1-\alpha_2)(1-\zeta)}} \log\left(\frac{4L_{\alpha,\beta,\zeta} \log(D_{\alpha,\beta,\zeta}/\Delta_S^*)^{\beta'}}{\Delta_S^*}\right)^\beta,
\end{aligned} \tag{57}$$

where $M = \sum_{k=1}^{\infty} k^\beta \cdot 2^{-k}$ is a constant only dependent on α, β, ζ, and the second to the third inequality is following the same pattern in Inequality (10). Then, combining the satisficing regret incurred by round i_0 (56) and after i_0 (57), the total satisficing regret is bounded by

$$\begin{aligned}
\mathbb{E}[\texttt{Regret}_S] \leq &\frac{8C_1 K^{\alpha_1+1}(2^{\beta+1}\alpha_2^{-\beta} M + (1-\alpha_2)^{-\beta})}{(1-\zeta)^\beta} \left(\frac{L_{\alpha,\beta,\zeta}}{\Delta_S^*}\right)^{\frac{\alpha_2}{(1-\alpha_2)(1-\zeta)}} \\
&\cdot \log\left(\frac{D_{\alpha,\beta,\zeta}}{\Delta_S^*}\right)^{\frac{\alpha_2 \beta'}{(1-\alpha_2)(1-\zeta)}} \log\left(\frac{2L_{\alpha,\beta,\zeta}}{\Delta_S^*} \log\left(\frac{D_{\alpha,\beta,\zeta}}{\Delta_S^*}\right)^{\beta'}\right)^\beta.
\end{aligned}$$

Furthermore, since by definition of $t_i = \lceil 2^{i/\alpha_2} \rceil$, the number of rounds that end within the time horizon is bounded by $2^{i/\alpha_2} \leq T$, *i.e.*, $i \leq \lfloor \alpha_2 \log_2(T) \rfloor$ because i is an integer. Then, the maximum number of rounds is bounded by $i_{\max} = \lceil \alpha_2 \log_2(T) \rceil$. Therefore, we have that the satisficing regret is bounded by

$$\begin{aligned}
\mathbb{E}[\texttt{Regret}_S] &\leq \sum_{i=1}^{i_{\max}} \frac{4C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} \gamma_i^{-\alpha_2/(1-\alpha_2)} \log(4/\gamma_i)^\beta \\
&\leq \sum_{i=1}^{i_{\max}} \frac{4C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} 2^{i+1} \log(8T)^\beta \\
&\leq \frac{32C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} T^{\alpha_2} \log(8T)^\beta
\end{aligned}$$

where the first inequality is based on Proposition 7, the second inequality is based on the definition $\gamma_i = 2^{-i(1-\alpha_2)/\alpha_2}$ and $4/\gamma_i \leq 8T$ as shown in (12).

H.3. Proof of Proposition 7

We follow the proof of Proposition 1 in Section A.1. Recall that $t_i = \lceil \gamma_i^{-1/(1-\alpha)} \rceil$, similar to (14), the satisficing regret incurred when running $\texttt{ALG}(t_i)$ is bounded by

$$\mathbb{E}\left[\sum_{s=1}^{t_i} \max\{S - r(A_s), 0\}\right] \le \mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right] \le \frac{2C_1 K^{\alpha_1}}{(1-\alpha_2)^\beta} \gamma_i^{-\alpha_2/(1-\alpha_2)} \log(4/\gamma_i)^\beta. \tag{58}$$

Since $\hat{A}_i$ is the most frequently selected arm, it has been pulled for at least t_i/K times during $\texttt{ALG}(t_i)$, and we further have

$$\mathbb{E}[\max\{S - r(\hat{A}_i), 0\}] \le \mathbb{E}[r(A^*) - r(\hat{A}_i)] \le \frac{\mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right]}{t_i/K} \le \frac{2C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} \gamma_i \log(4/\gamma_i)^\beta. \tag{59}$$

If $r(\hat{A}_i) \ge S$, then for every k, round i is not terminated after pulling arm $\hat{A}_i$ for k times (in addition to the number of pulls τ_i during $\texttt{ALG}(t_i)$) occurs with probability at most

$$\begin{aligned}
&\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau_i + k} - \sqrt{\frac{2k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{\tau_i + k}} > S\right) \\
\le &\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau_i + k} - \sqrt{\frac{2k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{\tau_i + k}} > r(\hat{A}_i)\right) \\
\le &\exp\left(-2k^\zeta + \log\left(\frac{\zeta}{8\Gamma(\zeta^{-1})}\right)\right) \le \frac{\zeta}{8\Gamma(\zeta^{-1})} \exp\left(-k^\zeta\right).
\end{aligned}$$

Therefore expected number of arm pulls of $\hat{A}_i$ after first pulling it τ_i times is bounded by

$$\begin{aligned}
&\mathbb{E}\Big[\text{Number of times } \hat{A}_i \text{ is pulled in round } i \text{ after the first } T_i \text{ times}\Big] \\
= &\sum_{s=1}^{\infty} \Pr(k \ge s) = \frac{\zeta}{8\Gamma(\zeta^{-1})} \sum_{s=1}^{\infty} \exp\left(-s^\zeta\right) \le \frac{\zeta}{8\Gamma(\zeta^{-1})} \int_0^\infty \exp\left(-x^\zeta\right) \mathrm{d}x \\
\le &\frac{\zeta}{8\Gamma(\zeta^{-1})} \cdot \zeta^{-1}\Gamma(\zeta^{-1}) = \frac{1}{8} \le 1.
\end{aligned}$$

Summing up the satisficing regret incurred in oracle running (58) and LCB tests (59), we have that the satisficing regret of round i is bounded by

$$\begin{aligned}
&\frac{2C_1 K^{\alpha_1}}{(1-\alpha_2)^\beta} \gamma_i^{-\alpha_2/(1-\alpha_2)} \log(4/\gamma_i)^\beta + 1 \cdot \frac{2C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} \gamma_i \log(4/\gamma_i)^\beta \\
\le &\frac{4C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} \gamma_i^{-\alpha_2/(1-\alpha_2)} \log(4/\gamma_i)^\beta.
\end{aligned}$$

H.4. Proof of Proposition 8

We follow the proof of Proposition 4 in Section E.1. Recall from the proof of Proposition 7, we have

$$\mathbb{E}[\max\{S - r(\hat{A}_i), 0\}] \le \mathbb{E}[r(A^*) - r(\hat{A}_i)] \le \frac{\mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right]}{t_i/K} \le \frac{2C_1 K^{\alpha_1+1}}{(1-\alpha_2)^\beta} \gamma_i \log(4/\gamma_i)^\beta.$$

According to the condition of this proposition, we have i satisfies the condition

$$\frac{32C_1K^{\alpha_1+1}(1-\zeta)^{-\zeta/2}}{(1-\alpha)^\beta}\sqrt{2+4\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}\gamma_i^{1-\zeta}\log(4/\gamma_i)^{\max\{\beta,1/2\}} \le \Delta_S^*.$$

Note that for any $0<\zeta<1$, $(1-\zeta)^{-\zeta/2}>1$ and $\gamma_i^{1-\zeta}>\gamma_i$. Therefore i satisfies

$$\frac{32C_1K^{\alpha_1+1}}{(1-\alpha)^\beta}\gamma_i\log(4/\gamma_i)^{\max\{\beta,1/2\}} \le \Delta_S^*,$$

by the Markov's inequality we have

$$\Pr(r(A^*)-r(\hat{A}_i)\ge\Delta_S^*/2) \le \Pr\left(r(A^*)-r(\hat{A}_i)\ge\frac{16C_1K^{\alpha_1+1}}{(1-\alpha)^\beta}\log(4/\gamma_i)^\beta\right)\le\frac{1}{8}. \tag{60}$$

Now, if $r(\hat{A}_i)\ge r(A^*)-\Delta_S^*/2$, which happens with probability at least $7/8$ according to the above inequality, then $r(\hat{A}_i)\ge S+\Delta_S^*/2$ holds. On the other hand, if i satisfies the condition stated in the proposition, then

$$(1-\zeta)^{-\zeta/2}\cdot\sqrt{2+4\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}\gamma_i^{1-\zeta}\log(4/\gamma_i)^{1/2}\le\Delta_S^*/16.$$

Note that for $f(y)=y^\zeta/(\tau_i+y)$, it reaches its maximum when

$$\frac{\mathrm{d}}{\mathrm{d}y}f(y)=\frac{\zeta y^{\zeta-1}\tau_i+y^\zeta(\zeta-1)}{(\tau_i+y)^2}=0.$$

Therefore the maximum of $f(y)$ is reached at $\zeta\tau_i/(1-\zeta)$. For simplicity, we set $D=\log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)$. Similar to Inequality (37), the confidence radius for any k is bounded by

$$\begin{aligned}
&\sqrt{\frac{2k^\zeta+D}{\tau_i+k}}\le\sqrt{\frac{(D+2)k^\zeta}{\tau_i+k}}\le\sqrt{\frac{(D+2)\left(\frac{\zeta\tau_i}{1-\zeta}\right)^\zeta}{\tau_i+\left(\frac{\zeta\tau_i}{1-\zeta}\right)}}\\
\le&\sqrt{\frac{(D+2)(1-\zeta)^{-\zeta}\cdot\zeta^\zeta\tau_i^\zeta}{\tau_i}}\le\sqrt{D+2}K^{(1-\zeta)/2}(1-\zeta)^{-\zeta/2}\gamma_i^{1-\zeta}\\
\le&2\sqrt{D+2}K(1-\zeta)^{-\zeta/2}\gamma_i^{1-\zeta}\log(4/\gamma_i)^{1/2}\le\frac{\Delta_S^*}{8},
\end{aligned}$$

where the only difference to Inequality (37) is the constant term, *i.e.*, $D+2$ instead of $D+1$, and $\tau_i\ge t_i/K\ge\gamma_i^{-1/(1-\alpha)}/K\ge\gamma_i^{-2}/K$ is used. Suppose round i is terminated for some k, then

$$\frac{\hat{r}_i^{\mathrm{tot}}}{\tau_i+k}-\sqrt{\frac{2k^\zeta+D}{\tau_i+k}}<S=r(A^*)-\Delta_S^*\le r(\hat{A}_i)-\frac{\Delta_S^*}{2}\le r(\hat{A}_i)-2\sqrt{\frac{2k^\zeta+D}{\tau_i+k}},$$

or equivalently

$$\frac{\hat{r}_i^{\mathrm{tot}}}{\tau_i+k}\le r(\hat{A}_i)-\sqrt{\frac{2k^\zeta+D}{\tau_i+k}}.$$

By the Hoeffding's inequality we have

$$\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau_i+k} \le r(\hat{A}_i) - \sqrt{\frac{2k^\zeta + D}{\tau_i+k}}\right) \le \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-2k^\zeta\right) \le \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^\zeta\right). \tag{61}$$

Therefore conditioned on $r(\hat{A}_i) \ge r(A^*) - \Delta_S^*/2$ the probability that round i is terminated by the end of the horizon is bounded by

$$\begin{aligned}&\sum_{k=1}^{\infty}\frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^\zeta\right) \le \frac{\zeta}{8\Gamma(\zeta^{-1})}\int_0^\infty \exp\left(-x^\zeta\right)\mathrm{d}x\\ \le&\frac{\zeta}{8\Gamma(\zeta^{-1})}\cdot\zeta^{-1}\Gamma(\zeta^{-1}) = \frac{1}{8}.\end{aligned} \tag{62}$$

Combining (60) and (62) we conclude that the probability that round i terminates by the end of the time horizon is upper bounded by $1/4$.

H.5. Proof of Theorem 11

We first prove that the regret incurred in round i is at most

$$\frac{2C_1K^{\alpha_1+1}}{(1-\alpha_2)^\beta}\gamma_i^{-\alpha_2/(1-\alpha_2)}\log(4/\gamma_i)^\beta.$$

Similar to (14) with modification on constant term C_1 to $C_1K^{\alpha_1}$, the standard regret incurred when running $\texttt{ALG}(t_i)$ is bounded by

$$\mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right] \le \frac{2C_1K^{\alpha_1}}{(1-\alpha_2)^\beta}\gamma_i^{-\alpha_2/(1-\alpha_2)}\log(4/\gamma_i)^\beta. \tag{63}$$

Since $\hat{A}_i$ is the most frequently selected arm, it has been pulled for at least t_i/K times during $\texttt{ALG}(t_i)$, and similar to (15) we further have

$$\mathbb{E}[r(A^*) - r(\hat{A}_i)] \le \frac{\mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s)\right]}{t_i/K} \le \frac{2C_1K^{\alpha_1+1}}{(1-\alpha_2)^\beta}\gamma_i\log(4/\gamma_i)^\beta. \tag{64}$$

Since $r(\hat{A}_i) \le r(A^*) < S$, then for every k, round i is not terminated after pulling arm $\hat{A}_i$ for k times (in addition to the number of pulls τ_i during $\texttt{ALG}(t_i)$) occurs with probability at most

$$\begin{aligned}&\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau_i+k} - \sqrt{\frac{2k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{\tau_i+k}} > S\right)\\ \le&\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau_i+k} - \sqrt{\frac{2k^\zeta + \log\left(8\zeta^{-1}\Gamma(\zeta^{-1})\right)}{\tau_i+k}} > r(\hat{A}_i)\right)\\ \le&\exp\left(-2k^\zeta + \log\left(\frac{\zeta}{8\Gamma(\zeta^{-1})}\right)\right) = \frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^\zeta\right).\end{aligned}$$

Therefore expected number of arm pulls of $\hat{A}_i$ after first pulling it τ_i times is bounded by

$$\begin{aligned}&\sum_{s=1}^{\infty}\frac{\zeta}{8\Gamma(\zeta^{-1})}\exp\left(-k^{\zeta}\right)\leq\frac{\zeta}{8\Gamma(\zeta^{-1})}\int_0^{\infty}\exp\left(-x^{\zeta}\right)\mathrm{d}x\\ \leq&\frac{\zeta}{8\Gamma(\zeta^{-1})}\cdot\zeta^{-1}\Gamma(\zeta^{-1})=\frac{1}{8}\leq 1.\end{aligned}\tag{65}$$

Summing up the standard regret in oracle running (63) and LCB tests as a combination of (64) and (65), we have that the standard regret of round i is bounded by

$$\frac{2C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}}\gamma_i^{-\alpha_2/(1-\alpha_2)}\log(4/\gamma_i)^{\beta}.$$

Since the length of the time horizon is T time steps, we have $t_i\leq T$, following $i_{\max}=\lceil\alpha_2\log_2(T)\rceil$. Therefore, following similar chain of inequality in (13), we have that the standard regret is bounded by

$$\begin{aligned}\mathbb{E}[\texttt{Regret}]\leq&\sum_{i=1}^{i_{\max}}\frac{4C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}}\gamma_i^{-\alpha_2/(1-\alpha_2)}\log(8T)^{\beta}\\ \leq&\sum_{i=1}^{i_{\max}}\frac{4C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}}2^{i+1}\log(8T)^{\beta}\\ \leq&\frac{16C_1K^{\alpha_1+1}}{(1-\alpha_2)^{\beta}}T^{\alpha_2}\log(8T)^{\beta}\end{aligned}$$

Therefore the regret is bounded by $C_1K^{\alpha_1+1}T^{\alpha_2}\cdot\mathrm{polylog}(T)$.

I. Choices of C_1,α,β for Several Bandit Optimization Problem Classes

In this section, we provide a detailed discussion on the choice of C_1,α,β for several bandit optimization problem classes. In three cases

I.1. Finite-armed Bandits

Consider any instance of finite-armed bandits with mean reward bounded by $[0,1]$. By Theorem 7.2 of Lattimore and Szepesvári (2020), for any given time horizon $t\geq 2$, the standard regret of the UCB algorithm is upper bounded by

$$8\sqrt{Kt\log(t)}+3\sum_{k=1}^{K}(r(k^*)-r(k))\leq 8\sqrt{Kt\log(t)}+3K,$$

where $K\geq 2$ is the number of arms and k^* is the optimal arm.

When $t>K$, the regret is upper bounded by $8\sqrt{Kt\log(t)}+3K\leq 11\sqrt{Kt\log(t)}$, where the last inequality is due to $t>K\geq 2$, thus $t\geq 3$ and $\log(t)\geq\log(3)>1$; When $t\leq K$, the regret is upper bounded by $t\leq\sqrt{Kt}\leq 11\sqrt{Kt\log(2)}\leq 11\sqrt{Kt\log(t)}$.

Thus for any given time horizon $t\geq 2$, the regret of the UCB algorithm is bounded by $11\sqrt{Kt\log(t)}$. Therefore, by setting $C_1=11\sqrt{K}$ and $\alpha=\beta=1/2$ and using UCB algorithm as the oracle, Condition 1 is satisfied for finite-armed bandits.

I.2. Concave Bandits

Consider any concave bandits with $\mathcal{A} \subseteq \mathbb{R}^d$ being a compact and convex set contained in a Euclidean ball with radius 1, and reward function being concave and 1-Lipschitz. By Theorem 2 of Agarwal et al. (2011), for any given time horizon $t \geq 2$, the standard regret of Algorithm 2 in Agarwal et al. (2011) is upper bounded by

$$768d^3\sqrt{t}\log(t)^2\left(\frac{2d^2\log(d)}{c_2^2}+1\right)\left(\frac{4d^7c_1}{c_2^3}+\frac{d(d+1)}{c_2}\right)\left(\frac{4c_1d^4}{c_2^2}+11\right),$$

where c_1, c_2 are algorithm hyperparameters that satisfies $c_1 \geq 64$, $c_2 \leq 1/32$. Therefore Condition 1 is satisfied for concave bandits by using Algorithm 2 in Agarwal et al. (2011) and setting $\alpha = 1/2$, $\beta = 2$ and

$$C_1 = 768d^3\left(\frac{2d^2\log(d)}{c_2^2}+1\right)\left(\frac{4d^7c_1}{c_2^3}+\frac{d(d+1)}{c_2}\right)\left(\frac{4c_1d^4}{c_2^2}+11\right),$$

Furthermore, if the instance of concave bandit is one-dimensional, *i.e.*, $\mathcal{A} = [0,1]$, by Theorem 1 of Agarwal et al. (2011), for any given time horizon $t \geq 2$, the standard regret of Algorithm 1 in Agarwal et al. (2011) is bounded by

$$108\sqrt{t\log(t)}\cdot\log_{4/3}(t) = 108/\log(4/3)\sqrt{t}\cdot\log(t)^{3/2}.$$

Therefore Condition 1 is also satisfied for one-dimensional concave bandits by using Algorithm 1 in Agarwal et al. (2011) and setting $C_1 = 108/\log(4/3)$, $\alpha = 1/2$, $\beta = 3/2$.

I.3. Lipschitz Bandits

Consider any Lipschitz bandits with arm set being $[0,1]^d$, reward function being L-Lipschitz and mean reward bounded on $[0,1]$ (here we define Lipschitz continuity in the sense of ∞-norm). By Section 2.2 of Bubeck et al. (2011), if the regret of the UCB algorithm for finite-armed bandits with K arms is bounded by $c_{\mathrm{ALG}}\sqrt{Kt\log(t)}$ for any $t \geq 2$, then for any given time horizon $t \geq 2$, the standard regret of the uniformly discretized UCB algorithm introduced by Bubeck et al. (2011) (hereafter referred to as the "Uniform UCB") is upper bounded by

$$(1+c_{\mathrm{ALG}})L^{d/(d+2)}t^{(d+1)/(d+2)}\cdot\sqrt{\log(t)}.$$

Using the results in Section I.1, we have $c_{\mathrm{ALG}} = 11$, thus the regret of the Uniform UCB algorithm is bounded by

$$12L^{d/(d+2)}t^{(d+1)/(d+2)}\cdot\sqrt{\log(t)}.$$

Therefore, by setting $C_1 = 12L^{d/(d+2)}$, $\alpha = (d+1)/(d+2)$, $\beta = 1/2$ and using Uniform UCB as the oracle, Condition 1 is satisfied for Lipschitz bandits.

J. Ablation Study for the Three Steps of SELECT

In this section, we use an ablation study to demonstrate that all three steps of SELECT play a vital role in obtaining a constant satisficing regret. Specifically, we compare the performance of SELECT with the following adapted versions of SELECT:

- SELECT **without Step 1:** instead of identifying candidate satisficing arms using a learning oracle, the candidate satisficing arm is drawn uniformly at random from the arm set before we proceed to forced sampling and LCB test in each round;
- SELECT **without Step 2:** After identifying a candidate satisficing arm from Step 1 in each round, the forced sampling in Step 2 is skipped and a LCB test on the candidate satisficing arm is immediately started;
- SELECT **without Step 3:** Each round is terminated after running Step 1 and Step 2 without entering the LCB test in Step 3.

Setup: We conduct the ablation test using an instance of the Lipschitz bandit. We use the reward function defined in Section 7.3, and set the satisficing threshold $S = 0.7$. The learning oracle we use in Step 1 is Uniform UCB introduced in Bubeck et al. (2011).

Results: The results are provided in Figure 10. One can see that after removing any of the three steps, our algorithm is unable to maintain constant satisficing regret in the realizable case. Specifically, Step 1 serves the purpose of efficiently identifying a candidate satisficing arm. If Step 1 is replaced by randomly drawing candidate satisficing arms from the arm set, then the satisficing regret is almost linear in T, indicating that the algorithm struggles to identify satisficing arms without the help of the learning oracle. Step 2 and Step 3 combined mainly serve the purpose of exploiting the candidate satisficing arm and at the same time, testing whether it is truly satisficing. Without either Step 2 or Step 3, the satisficing regret starts a round of rapid increase whenever the time horizon T exceeds certain thresholds, thus fail to maintain constant satisficing regret. This indicates that without either Step 2 or Step 3, the algorithm almost always terminates a round shortly after completing Step 1 even if the candidate satisficing arm obtained from Step 1 is indeed satisficing. We refer to Remark 4 for a detailed discussion on the reason for this. Therefore we conclude that all three steps of SELECT play a vital role in obtaining a constant satisficing regret bound in the realizable case.

K. Robustness on the choice of γ_i

In Algorithm 1, we define $\gamma_i = 2^{-i(1-\alpha)/\alpha}$ for all $i \in \mathbb{Z}_+$, but all our results still hold if all γ_i are multiplied by a constant. In this section, we conduct numerical experiments on testing the robustness of SELECT in the choice γ_i.

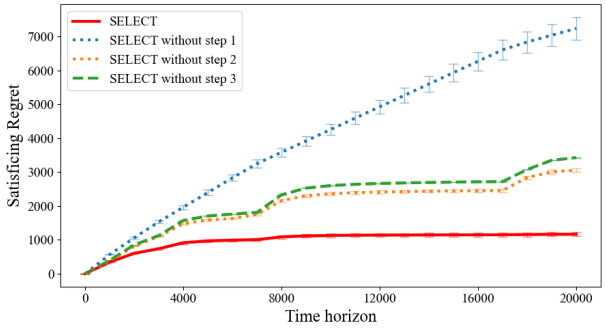
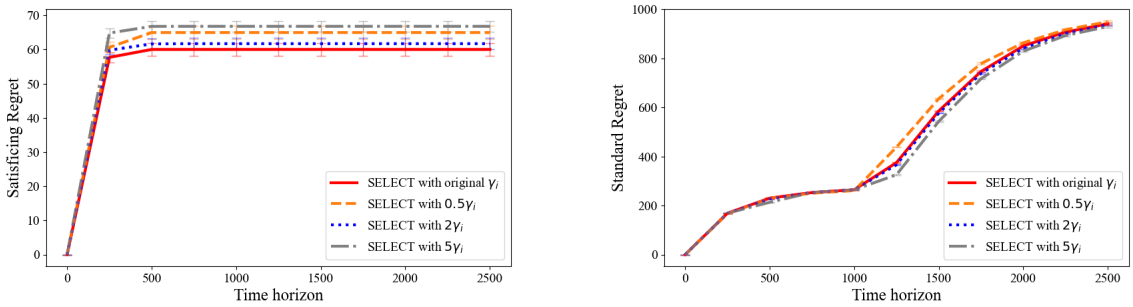


Figure 10 **Results of the ablation test**



(a) Realizable case (b) Non-realizable case

Figure 11 **Performance of `SELECT` with different choice of γ_i**

Setup: We conduct the experiment on the same instance of Lipschitz bandit in Section 7.3. We multiply the $\{\gamma_i\}$ in Algorithm 1 with a constant λ, and vary $\lambda = 0.5, 1, 2, 5$.

Results: The result of the realizable case is provided in Figure 11(a), and the result of the non-realizable case is provided in Figure 11(b). One can see from the results that different choices of γ_i have limited impact on the empirical performance of `SELECT`, both in the realizable and non-realizable case. Therefore the empirical performance of `SELECT` is robust in the choice of γ_i.

L. Ablation Test for Oracle Choice in Finite-Armed Bandits

In this section, we conduct an ablation study to demonstrate that in finite-armed bandits, using oracle algorithm with sub-linear standard regret bound instead of a uniform sampling, the approach used in Michel et al. (2023) and Garivier et al. (2019), plays a vital role in obtaining a small constant satisficing regret in the realizable case and a sub-linear standard regret in the non-realizable case.

The same as Section 7.1 and Section N, we consider an instance of finite-armed bandits with 4 arms, and the expected rewards of all arms are $\{0.2, 0.4, 0.6, 0.8\}$. We vary the length of the time horizon from 500 to 7000 with a stepsize of 500. We conduct experiments for both the realizable case and the non-realizable case. For the realizable case, we set the satisficing level $S \in \{0.65, 0.7, 0.75\}$;

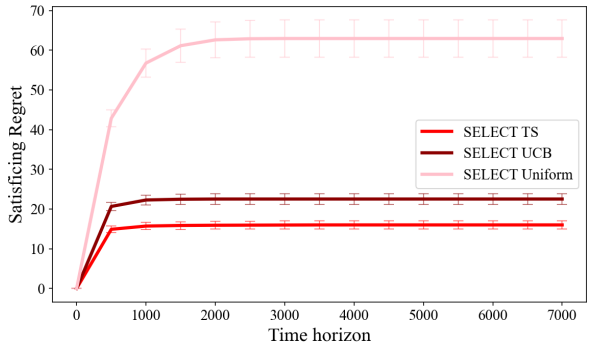
for the non-realizable case, we set the satisficing level $S = 1.5$. In both cases, we test the performance of SELECT-BoBW with different oracle algorithms, including Thompson sampling, UCB algorithm, and uniform exploration (*i.e.*, draw an arm uniformly at random at each time step). The experiment is repeated for 5000 times and we report the average satisficing regret in the realizable case and the average standard regret in the non-realizable case.

The results of the realizable case are presented in Figure 12(a), 12(b), and 12(c), and the results of the non-realizable case are presented in Figure 12(d). from Figure 12(a) and 12(b), one can observe that although SELECT-BoBW using uniform sampling as oracle algorithm is able to attain a constant expected satisficing regret, it exhibits a much slower convergence rate and incurs a much larger expected satisficing regret, compared to SELECT-BoBW using Thompson sampling or UCB as oracle algorithms. Furthermore, one can observe from Figure 12(c) that if S is large and Δ_S^* is small, SELECT-BoBW using uniform sampling as oracle algorithm might even fail to reach a constant satisficing regret within a time horizon of 7000. Furthermore, in all these three realizable cases, SELECT-BoBW using uniform sampling as oracle algorithm has higher variance according to the error bar. This is because unlike an oracle algorithm with sub-linear standard regret bound, uniform exploration does not converge to the optimal arm, and it is more likely that a non-satisficing arm is selected as a candidate for Step 2 and 3 in each round. When comparing SELECT-BoBW using Thompson sampling with SELECT-BoBW using UCB agorithm, one can observe that the performance of SELECT-BoBW with Thompson sampling outperforms SELECT-BoBW with UCB. This observation indicates that while any learning oracle with sub-linear expected standard regret enables SELECT-BoBW to attain a constant expected satisficing regret, SELECT-BoBW can also attain better empirical performance if a learning oracle with better empirical performance is implemented in Step 1.

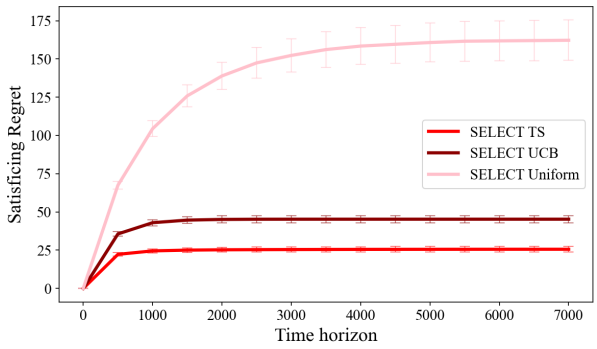
From Figure 12(d), one can see that in the non-realizable case, SELECT-BoBW using uniform exploration as oracle algorithm incurs linear expected standard regret in T, and the expected standard regret is much larger than that of SELECT-BoBW using sub-linear oracle algorithms. Therefore we conclude that using an oracle algorithm with sub-linear standard regret bound instead of a uniform exploration is necessary.

M. Ablation Test for Design of $\{\gamma_i\}$

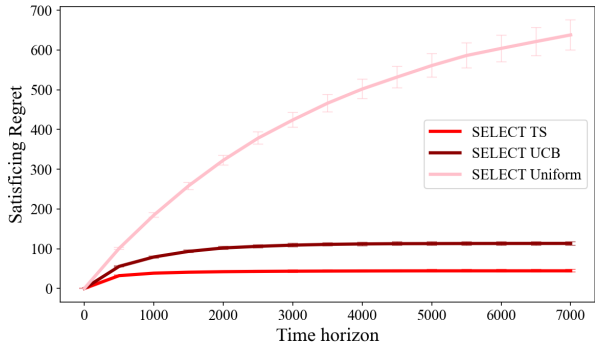
In order to highlight the importance of exponential decay of $\{\gamma_i\}$, in this section, We conduct an ablation test comparing different designs of $\{\gamma_i\}$ using an instance of Lipschitz bandit. Specifically, we use the instance of Lipschitz bandit used in Section 7.3, where the arm set is $\mathcal{A} = [0, 1]^2$, and the reward function is $r(x, y) = \min\{1, e^{-100((x-0.5)^2+(y-0.6)^2)}\}$. We vary the length of the time horizon from 500 to 10000 with a stepsize of 500. We use the Uniform UCB as the learning oracle of SELECT,



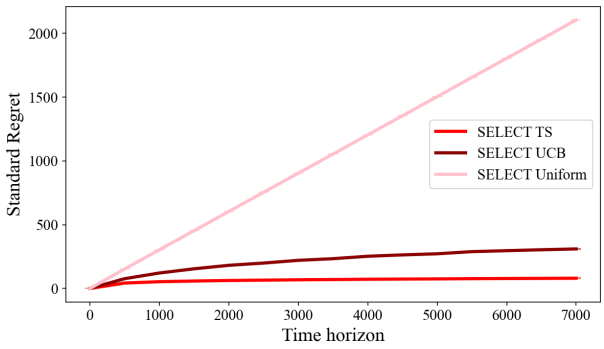
(a) Satisficing Regret in the Realizable Case $S = 0.65$



(b) Satisficing Regret in the Realizable Case $S = 0.7$



(c) Satisficing Regret in the Realizable Case $S = 0.75$



(d) Standard Regret in the Non-Realizable Case

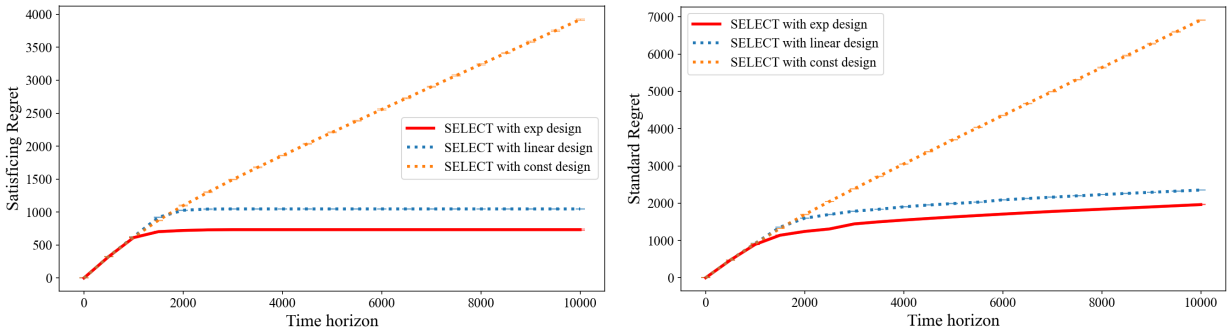
Figure 12 Performance of `SELECT-BoBW` using Thompson sampling, UCB, and uniform sampling as oracle algorithm

and we compare `SELECT` equipped with the following three designs of $\{\gamma_i\}$: 1) exponential decay, $\gamma_i = 2^{-i(1-\alpha)/\alpha}$, 2) linear decay $\gamma_i = (i/2)^{-1}$, and 3) constant $\gamma_i = 1/4$.

The result of the realizable case is provided in Figure 13(a), and the result of the non-realizable case is provided in Figure 13(b). One can see from Figure 13(a) that in the realizable case, although both exponentially decaying $\{\gamma_i\}$ and linearly decaying $\{\gamma_i\}$ attain a constant expected satisficing regret, the regret incurred by the exponentially decaying $\{\gamma_i\}$ is smaller than that incurred by linearly decaying $\{\gamma_i\}$. On the other hand, the satisficing regret with constant $\{\gamma_i\}$ increases linearly, failing to reach a constant value. In the non-realizable case, the exponential design of γ_i also incurs the least standard regret, followed by the linear design, while the constant design performs the worst. This shows that using a decreasing sequence of $\{\gamma_i\}$ is necessary for attaining a constant expected satisficing regret, and exponentially decaying $\{\gamma_i\}$ used in our algorithm is better than linearly decaying $\{\gamma_i\}$.

N. Best-of-Both-Worlds Regret Bound for Finite-Arm Bandits

For satisficing regret minimization in finite-arm bandits, existing works such as Garivier et al. (2019) establishes a constant expected satisficing regret bound of $\tilde{O}(K/\Delta_S)$, where Δ_S is the satisficing gap



(a) Realizable case

(b) Non-realizable case

Figure 13 **Performance of `SELECT` with different design of γ_i**

given by $\Delta_S = \min\{S - r(A) : A \in \mathcal{A}, r(A) < S\}$. On the other hand, our algorithm `SELECT` is able to attain a $\tilde{O}(K/\Delta_S^*)$ expected regret bound, where Δ_S^* is the exceeding gap defined as $\Delta_S^* = r(A^*) - S$. When comparing the our expected regret bound in Corollary 1 with the one in Garivier et al. (2019), one can see that our algorithm `SELECT` potentially has a better performance when Δ_S^* is large and Δ_S is small, while the algorithm in Garivier et al. (2019) potentially has a better performance when Δ_S^* is small and Δ_S is large.

In order to establish an algorithm that attains a "best of both worlds" and works well in both cases, in this section, we present `SELECT-BoBW`, an adapted version of `SELECT` for finite-armed bandits that incurs an expected satisficing regret bound $\tilde{O}(\min\{1/\Delta_S, 1/\Delta_S^*\})$. Specifically, the expected satisficing regret bound of `SELECT-BoBW` achieves the "best-of-both-worlds", dominating both regret bound of `SELECT` and the one in Garivier et al. (2019). The complete algorithm is provided in Algorithm 4.

In `SELECT-BoBW`, we can use either the UCB algorithm (see, *e.g.*, Bubeck and Cesa-Bianchi (2012)) or Thompson sampling (see, *e.g.*, Agrawal and Goyal (2017)) as the learning oracle `ALG`.

We use the following theorem to establish an expected satisficing regret bound for `SELECT-BoBW`.

Theorem 12. *If $r(A^*) \geq S$, then the expected satisficing regret of `SELECT-BoBW` is bounded by*

$$\mathbb{E}[\textit{Regret}_S] \leq \min\left\{\frac{K}{\Delta_S^*} \cdot \text{polylog}\left(\frac{K}{\Delta_S^*}\right), \frac{K}{\Delta_S} \cdot \text{polylog}\left(\frac{K}{\Delta_S^* \Delta_S}\right), \sqrt{KT} \cdot \text{polylog}(T)\right\}. \tag{66}$$

Proof of Theorem 12 is provided in Section N.2.

We also show that in the non-realizable case, `SELECT-BoBW` enjoys the same standard regret bound as the oracle shown in Theorem 13.

Theorem 13. *If $r(A^*) < S$, the standard regret of `SELECT-BoBW` is bounded by*

$$\mathbb{E}[\textit{Regret}] \leq \sqrt{KT} \cdot \text{polylog}(T).$$

Proof of Theorem 13 is provided in Section N.5 of the appendix.

Algorithm 4 Satisficing Regret Minimization for Finite-Armed Bandits (`SELECT-BoBW`)

Input: time horizon T, satisficing level S
Set $\gamma_i \leftarrow 2^{-i}$ for all $i = 1, 2, 3 \dots$
for round $i = 1, 2, \dots$ **do**
  Set $t_i \leftarrow \lceil 4^i \rceil$
  Run `ALG`(t_i), let $\{\bar{A}_s\}_{s \in [t_i]}$ be the trajectory of arm pulls
  Sample q uniformly at random from $\{1, 2, \dots, t_i\}$ and set $\hat{A}_i \leftarrow \bar{A}_q$
  Set $T_i \leftarrow 4^i$, $k \leftarrow 0$, $\hat{r}_i^{\text{tot}} \leftarrow 0$
  for $\tau = 1, 2, \cdots, T_i$ **do**
    Pull arm $\hat{A}_i$ and observe Y_t, where t is the current time step
    Update $\hat{r}_i^{\text{tot}} \leftarrow \hat{r}_i^{\text{tot}} + Y_t$
    Set

$$UCB(\hat{A}_i) \leftarrow \frac{\hat{r}_i^{\text{tot}}}{\tau} + \sqrt{\frac{4\log(\tau+1)}{\tau}}$$

    if $UCB(\hat{A}_i) < S$ **then**
      Continue to next round $i+1$
    end if
  end for
  while $LCB(\hat{A}_i) \geq S$ **do**
    Set $k \leftarrow k+1$
    Pull arm $\hat{A}_i$ and observe Y_t, where t is the current time step
    Update $\hat{r}_i^{\text{tot}} \leftarrow \hat{r}_i^{\text{tot}} + Y_t$
    Set

$$LCB(\hat{A}_i) \leftarrow \frac{\hat{r}_i^{\text{tot}}}{T_i + k} - \sqrt{\frac{4\log(T_i+k)}{T_i+k}}$$

  end while
end for

Remark 11. *We would like to remark that forced sampling is not necessary in attaining a constant expected satisficing regret in finite-armed bandits. In fact, a more simplified version of* `SELECT-BoBW` *is to directly use the arm that is pulled most frequently along the trajectory during* `ALG`(t_i) *as the candidate arm, similar to Algorithm 3, and leverage the observed reward of this arm during oracle running. Using similar arguments in the proof of Theorem 12, we can prove that the expected satisficing regret of* `SELECT-Arm` *without forced sampling in the realizable case is bounded by*

$$\mathbb{E}[\texttt{Regret}_S] \leq \min\left\{\min\left\{\frac{K}{\Delta_S^*}, \frac{K}{\Delta_S}\right\} \cdot \text{polylog}\left(\frac{K}{\Delta_S^*}\right), K\sqrt{T} \cdot \text{polylog}(T)\right\}. \tag{67}$$

In other words, after removing forced sampling, the algorithm is still able to attain a constant expected satisficing regret bound of $\tilde{O}(\min\{1/\Delta_S, 1/\Delta_S^*\})$ *similar to* `SELECT-BoBW`*. However, when* Δ_S^* *is small, the algorithm without forced sampling potentially incurs a larger expected satisficing regret of* $\tilde{O}(K\sqrt{T})$ *rather than the corresponding* $\tilde{O}(\sqrt{KT})$ *expected satisficing regret bound for* `SELECT-BoBW`*. This is because the most frequently pulled arm is pulled for at least* t_i/K *times during* `ALG`(t_i)*, and its (satisficing) regret of a single pull is bounded by* $\tilde{O}(t_i^\alpha/(t_i/K)) = \tilde{O}(K\gamma_i)$*, which*

only suffers from a difference in constant term K compared to the original one with forced sampling. At the same time, its width of the confidence interval at the beginning of LCB test is of order $\tilde{O}((t_i/K)^{-1/2}) = \tilde{O}(\gamma_i\sqrt{K})$ by leveraging its observed reward during $\texttt{ALG}(t_i)$, which still only suffers from a difference in constant term $\sqrt{K}$ compared to the original one. That is, both the expected satisficing regret incurred in round i given in Proposition 9, and the round i_0 that we can start to bound the exit probability given by Proposition 10, only change at most by a constant term K. On the other hand, by skipping the forced sampling, we are able to get rid of the $1/\Delta_S \cdot \log(1/\Delta_S)$ in Proposition 9 which is incurred in forced sampling. Finally, we are still able to bound the expected satisficing regret by the Best-of-Both-Worlds $\tilde{O}(\min\{K/\Delta_S^*, K/\Delta_S\})$ and a slightly worse sub-linear satisficing regret $\tilde{O}(K\sqrt{T})$.

N.1. Numerical Experiment

To test the performance of `SELECT-BoBW`, we conduct a numerical experiment under a wide range of (Δ_S, Δ_S^*) pairs on finite-armed bandits. We consider an instance of 4 arms, and the expected rewards of all arms are $\{0.2, 0.4, 0.6, 0.8\}$, the same as the setting in Section 7.1. We fix the time horizon to be 10000, and vary the satisficing level S from 0.62 to 0.78 with a stepsize of 0.01, resulting in the pair $(\Delta_S, \Delta_S^*) \in \{(0.02, 0.18), (0.04, 0.16), \cdots, (0.18, 0.02)\}$. We compare `SELECT-BoBW` against SAT-UCB and SAT-UCB+ in Michel et al. (2023), and the "Phase Exploration algorithm" in Garivier et al. (2019) with μ^* changed to S. For `SELECT-BoBW`, we use its simplified version noted in Remark 11, and use UCB algorithm and Thompson sampling respectively as the learning oracle. The experiment is repeated for 1000 times and we report the average satisficing regret.

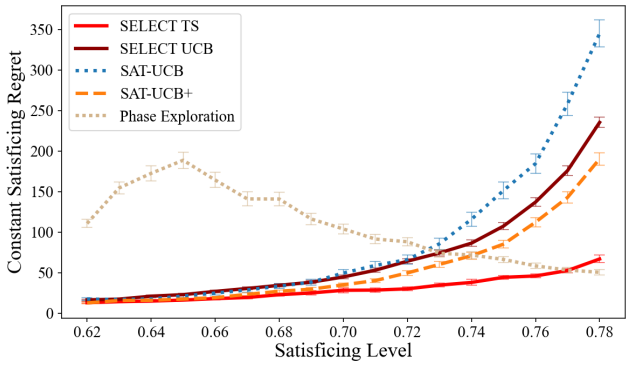


Figure 14 **Constant satisficing regret changing with satisficing level S at time horizon $T = 10000$**

The results are presented in Figure 14. One can see that `SELECT-BoBW` using Thompson sampling as oracle algorithm has the lowest constant expected satisficing regret under most of the (Δ_S, Δ_S^*) pairs, except for $(\Delta_S, \Delta_S^*) = (0.18, 0.02)$ (*i.e.*, $S = 0.78$) where the Phase Exploration algorithm has a lower satisficing regret due to an additional polylog$(1/\Delta_S^*)$ in the satisficing regret bound

of SELECT-BoBW. The constant expected satisficing regret of SELECT-BoBW using UCB as oracle algorithm is comparable to that of the heuristic algorithm SAT-UCB+, and is lower than that of SAT-UCB. Note that for the Phase Exploration algorithm, it fails to converge to a constant satisficing regret when $S < 0.67$, even if we increase the time horizon to $T = 20000$. Therefore the reported satisficing regret of the Phase Exploration algorithm when $S < 0.67$ is smaller than its constant expected satisficing regret.

N.2. Proof of Theorem 12

For any sub-optimal arm $a \in \mathcal{A}$, we define $\Delta_a = r(A^*) - r(a)$, and $N_a(t)$ as the number of times arm a is pulled if an algorithm (either the UCB algorithm or Thompson sampling) is run over a time horizon of t time periods. By Theorem 2.1 of Bubeck and Cesa-Bianchi (2012) and Theorem 1.1 of Agrawal and Goyal (2017), there exists constants $c_1, c_2 > 0$ such that both the UCB algorithm and Thompson sampling satisfies the following two properties:

1. **Worst-case expected standard regret bound:** the expected standard regret is upper bounded by $c_1\sqrt{KT\log(T)}$;
2. **Instance-dependent bound on expected arm pulls:** The expected number of arm pull of any sub-optimal arm a is upper bounded by $\mathbb{E}[N_a(t)] \le c_2 \log(t)/\Delta_a^2$.

Proof of the theorem relies on two critical results. First, we show an upper bound on the satisficing regret of round i.

Proposition 9. *If $r(A^*) \ge S$, then the expected satisficing regret incurred in round i is bounded by*

$$\min\left\{4c_1\sqrt{K}\gamma_i^{-1}\log\left(\frac{1}{\gamma_i}\right)^{1/2}, \frac{66}{\Delta_S}\log\left(\frac{5}{\Delta_S}\right) + \frac{K}{\Delta_S}\cdot 2c_2\log\left(\frac{1}{\gamma_i}\right)\right\}.$$

Proof of Proposition 9 is provided in Section N.3 of the appendix.

Next, we show that once SELECT-BoBW runs for enough rounds, it is unlikely to start a new round.

Proposition 10. *If $r(A^*) \ge S$, then for every i that satisfies*

$$16\sqrt{2}c_1\sqrt{K}\gamma_i\log(1/\gamma_i)^{1/2} \le \Delta_S^*,$$

the probability that round i ends within the time horizon T (conditioned on the event that round i is started within the time horizon) is upper bounded by $1/4$.

Proof of Proposition 10 is provided in Section N.4 of the appendix.

We define i_0 as

$$i_0 = \max\{i \in \mathbb{Z}_+ : 16\sqrt{2}c_1\sqrt{K}\gamma_i\log(1/\gamma_i)^{1/2} > \Delta_S^*\},$$

and denote $\delta_0 \in (0,1)$ the smallest solution to the following equation

$$\delta_0\log(1/\delta_0)^{1/2} = \frac{\Delta_S^*}{16\sqrt{2}c_1\sqrt{K}}.$$

Then we have

$$\frac{1}{\delta_0} = \frac{16\sqrt{2}c_1\sqrt{K}\log(1/\delta_0)^{1/2}}{\Delta_S^*} \leq \frac{16\sqrt{2}c_1\sqrt{K}\cdot 1/\delta_0^{1/2}}{\Delta_S^*} = \frac{16\sqrt{2}c_1\sqrt{K}}{\Delta_S^*}\cdot\frac{1}{\delta_0^{1/2}}.$$

Here the inequality is because $\log(x) < x$ for all $x > 0$. Therefore we have

$$\frac{1}{\delta_0} \leq \left(\frac{16\sqrt{2}c_1\sqrt{K}}{\Delta_S^*}\right)^2. \tag{68}$$

Therefore we have

$$\delta_0 = \frac{\Delta_S^*}{16\sqrt{2}c_1\sqrt{K}\cdot\log(1/\delta_0)^{1/2}} \geq \frac{\Delta_S^*}{32c_1\sqrt{K}\cdot\log(16\sqrt{2}c_1\sqrt{K}/\Delta_S^*)^{1/2}}.$$

Here the inequality follows from (68). By definition of i_0 we have that i_0 is the largest integer that satisfies $\gamma_{i_0} \geq \delta_0$. Therefore we get

$$\gamma_{i_0} \geq \frac{\Delta_S^*}{32c_1\sqrt{K}\cdot\log(16\sqrt{2}c_1\sqrt{K}/\Delta_S^*)^{1/2}}.$$

Recall that $\gamma_i = 2^{-i}$, the above inequality implies that

$$i_0 \leq \log_2\left(\frac{32c_1\sqrt{K}\cdot\log(16\sqrt{2}c_1\sqrt{K}/\Delta_S^*)^{1/2}}{\Delta_S^*}\right). \tag{69}$$

The satisficing regret bound depending only on $1/\Delta_S^*$ is exactly the same as the one in Theorem 1. In what follows, we focus on proving the expected satisficing regret bound that depends on Δ_S.

Using the result of Proposition 9, we have that the satisficing regret by round i_0 is bounded by

$$\begin{aligned}
&\sum_{i=1}^{i_0}\left(\frac{66}{\Delta_S}\log\left(\frac{5}{\Delta_S}\right) + \frac{K}{\Delta_S}\cdot 2c_2\log\left(\frac{1}{\gamma_i}\right)\right)\\
\leq& i_0\cdot\left(\frac{66}{\Delta_S}\log\left(\frac{5}{\Delta_S}\right) + \frac{K}{\Delta_S}\cdot 2c_2\log\left(\frac{1}{\delta_0}\right)\right)\\
\leq& \frac{1}{\Delta_S}\log_2\left(\frac{64c_1\sqrt{K}\cdot\log(128\sqrt{2}c_1\sqrt{K}/\Delta_S^*)^{1/2}}{\Delta_S^*}\right)\cdot\left(66\log\left(\frac{5}{\Delta_S}\right) + 4c_2K\log\left(\frac{16\sqrt{2}c_1\sqrt{K}}{\Delta_S^*}\right)\right)\\
\leq& \frac{K}{\Delta_S}\cdot\mathrm{polylog}\left(\frac{K}{\Delta_S\Delta_S^*}\right).
\end{aligned} \tag{70}$$

where the first inequality is by plugging into the definition of γ_i and $\gamma_i \geq \gamma_{i_0} \geq \delta_0$ for any $i \leq i_0$, and the second inequality is due to (68) and (69).

By the results of Proposition 10, for every $k \geq 1$, the probability that round $i_0 + k$ starts within the time horizon is at most 4^{-k+1}. Therefore the regret incurred after round i_0 is bounded by

$$\begin{aligned}
&\sum_{k=1}^{\infty} \frac{1}{\Delta_S} \left(66 \log\left(\frac{5}{\Delta_S}\right) + 2Kc_2 \log(1/\gamma_{i_0+k}) \right) \cdot 4^{-k+1} \\
\leq &\sum_{k=1}^{\infty} \frac{66}{\Delta_S} \log\left(\frac{5}{\Delta_S}\right) 4^{-k+1} + \sum_{k=1}^{\infty} \left(\log(4/\gamma_{i_0}) + k\right) \frac{2Kc_2}{\Delta_S} \cdot 4^{-k+1} \\
\leq &\frac{88}{\Delta_S} \log\left(\frac{5}{\Delta_S}\right) + \sum_{k=1}^{\infty} 2\log(4/\gamma_{i_0}) k \cdot 4^{-k+1} \cdot \frac{2Kc_2}{\Delta_S} \\
\leq &\frac{88}{\Delta_S} \log\left(\frac{5}{\Delta_S}\right) + \log(4/\delta_0) \frac{4Kc_2}{\Delta_S} \sum_{k=1}^{\infty} k \cdot 4^{-k} \\
\leq &\frac{88}{\Delta_S} \log\left(\frac{5}{\Delta_S}\right) + \frac{16Kc_2}{9} \log\left(\frac{L_{\alpha,\beta} c_1 \sqrt{K}}{\Delta_S^*}\right) \frac{1}{\Delta_S} \\
\leq &\frac{K}{\Delta_S} \cdot \mathrm{polylog}\left(\frac{K}{\Delta_S \Delta_S^*}\right).
\end{aligned} \tag{71}$$

The first inequality follows directly from the definition of $\gamma_i = 2^{-i}$ then $\log(1/\gamma_{i_0+k}) = \log(1/2^{-(i_0+k)}) = \log_2(2^k)/\log_2(e) + \log(1/\gamma_{i_0}) \leq \log(4/\gamma_{i_0}) + k$. The second inequality is due to $\log(4/\gamma_{i_0}) > 1 > 1/2$ for $\gamma_{i_0} < 1$, and for any $a, b \geq 1$, we have $2ab \geq ab + 1 \geq a + b$, and $\sum_{k\geq 1} 4^{-k+1} \leq 4/3$. The third inequality is similar to the case in satisficing regret by round i_0. The fourth inequality is plugging into Inequality (68) and $\sum_{k\geq 1} k \cdot 4^{-k} = 4/9$.

We note that $\sum 1/\Delta_S^a \leq K/(\min \Delta_S^a) = K/\Delta_S$. Taking the sum of (70) and (71), we have that the total expected satisficing regret is bounded by

$$\mathbb{E}[\texttt{Regret}_S] \leq \frac{K}{\Delta_S} \cdot \mathrm{polylog}\left(\frac{K}{\Delta_S \Delta_S^*}\right) \tag{72}$$

Following similar steps in the proof of Theorem 1, we can prove that the expected satisficing regret is upper bounded by

$$\mathbb{E}[\texttt{Regret}_S] \leq \min\left\{ \frac{K}{\Delta_S^*} \cdot \mathrm{polylog}\left(\frac{K}{\Delta_S^*}\right), \sqrt{KT} \cdot \mathrm{polylog}(T) \right\}. \tag{73}$$

Combining the regret bound in (72), (73), we have completed the proof of Theorem 12.

N.3. Proof of Proposition 9

Recall that $t_i = \gamma_i^{-2}$, then $\log(t_i) = \log(\gamma_i^{-2}) \leq 2\log(1/\gamma_i)$ and $t_i^{1/2} = 2^i = 1/\gamma_i$. Therefore in the realizable case, by the worst-case expected standard regret bound of $\texttt{ALG}$, the satisficing regret incurred when running $\texttt{ALG}(t_i)$ can be bounded by

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{t_i} \max\{S - r(A_s), 0\} \right] \\
\leq &\mathbb{E}\left[t_i r(A^*) - \sum_{s=1}^{t_i} r(A_s) \right] \leq c_1 \sqrt{K} t_i^{1/2} \log(t_i)^{1/2} \\
= &\sqrt{2} c_1 \sqrt{K} \gamma_i^{-1} \log(1/\gamma_i)^{1/2},
\end{aligned} \tag{74}$$

and the instance-dependent bound

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=1}^{t_i}\max\{S-r(A_s),0\}\right]\\
=&\mathbb{E}\left[\sum_{s=1}^{t_i}(S-r(A_s))\mathbf{1}(r(A_s)<S)\right]\\
\overset{\text{(i)}}{\leq}&\mathbb{E}\left[\sum_{a\in\mathcal{A},\, r(a)<S} N_a(t)(r(A^*)-r(a))\right]\\
\overset{\text{(ii)}}{\leq}&\sum_{a\in\mathcal{A},\, r(a)<S}\frac{c_2\log(t_i)}{\Delta_a}\overset{\text{(iii)}}{\leq}\sum_{a\in\mathcal{A},\, r(a)<S}\frac{c_2\log(t_i)}{\Delta_S^a}\\
=&\sum_{a\in\mathcal{A},\, r(a)<S}\frac{2c_2\log(1/\gamma_i)}{\Delta_S^a},
\end{aligned}\tag{75}$$

where step (i) is by decomposing satisficing regret by arm pulls, step (ii) follows from the instance-dependent bound on expected arm pulls for `ALG`, and inequality (iii) is because $\Delta_a = r(A^*) - r(a) > S - r(a) = \Delta_S^a$. Combining (74) and (75), we conclude that the satisficing regret incurred when running $\texttt{ALG}(t_i)$ is bounded by the

$$\mathbb{E}\left[\sum_{s=1}^{t_i}\max\{S-r(A_s),0\}\right]\leq\min\left\{\sqrt{2}c_1\sqrt{K}\gamma_i^{-1}\log(1/\gamma_i)^{1/2},\sum_{a\in\mathcal{A},\, r(a)<S}\frac{2c_2\log(1/\gamma_i)}{\Delta_S^a}\right\}\tag{76}$$

Then, we consider the expected satisficing regret incurred in the forced sampling with UCB test. Using similar arguments as in (16), the expected satisficing regret incurred in Step 2 is upper bounded by

$$T_i\cdot\frac{1}{t_i}\cdot c_1\sqrt{Kt_i\log(t_i)}=\sqrt{2}c_1\sqrt{K}\gamma_i^{-1}\log(1/\gamma_i)^{1/2}.\tag{77}$$

Then, we establish an expected satisficing regret bound in Step 2 with respect to Δ_S. If $r(\hat{A}_i)<S$, for simplicity, we denote $\hat{A}_i$ as a, and define τ_0 as

$$\tau_0=\max\left\{\tau:\,\sqrt{4\log(\tau+1)/\tau}>\Delta_S^a/2\right\}.$$

We first provide an upper bound on x given $x/\log(x+1)<y$. Because $\sqrt{x}>\log(x+1)$ for any $x\geq 1$, we know $y>x/\log(x+1)>\sqrt{x}$. Then, for any $x\geq 1$

$$x<y\log(x+1)<y\log(2x)<y\log(2y^2).$$

Plugging into the definition of τ_0 as $\tau/\log(\tau+1)<16/{\Delta_S^a}^2$, we get

$$\tau_0<\frac{16}{{\Delta_S^a}^2}\log\left(\frac{2\cdot 16^2}{{\Delta_S^a}^4}\right)<\frac{64}{{\Delta_S^a}^2}\log\left(\frac{5}{\Delta_S^a}\right).$$

For $\tau \le \tau_0$, the satisficing regret is bounded by

$$\tau_0 \cdot \Delta_S^a \le \frac{64}{\Delta_S^a} \log\left(\frac{5}{\Delta_S^a}\right). \tag{78}$$

For $\tau > \tau_0$, we know the confidence radius is less than $\Delta_S^a/2$, and the non-satisficing arm is pulled if its UCB is greater than S as

$$\frac{\hat{r}_i^{\text{tot}}}{\tau} + \sqrt{\frac{4\log(\tau)}{\tau}} \ge S = r(\hat{A}_i) + \Delta_S^a > r(\hat{A}_i) + 2\sqrt{\frac{4\log(\tau)}{\tau}},$$

this happens with probability

$$\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau} - \sqrt{\frac{4\log(\tau+1)}{\tau}} > S\right) \le \exp\left(-8\log(\tau+1)\right) = \frac{1}{(\tau+1)^8}.$$

Therefore the expected number of arm pulls of the non-satisficing $\hat{A}_i$ after first pulling it τ_0 times is bounded by

$$\begin{aligned} &\mathbb{E}\left[\text{Number of times } \hat{A}_i \text{ is pulled in forced sampling after the first } \tau_0 \text{ times}\right] \\ =&\sum_{s=\tau_0+1}^{\infty} \Pr(\tau \ge s) = \sum_{s=\tau_0+1}^{\infty} \frac{1}{(s+1)^8} \le 1. \end{aligned} \tag{79}$$

Combining the instance-dependent regret bound before τ_0 (78) and after τ_0 (79), recall that $\Delta_S^a < 1 < 1/\Delta_S^a$, we get

$$1 \cdot \Delta_S^a + \frac{64}{\Delta_S^a} \log\left(\frac{5}{\Delta_S^a}\right) \le \frac{1}{\Delta_S^a} + \frac{64}{\Delta_S^a} \log\left(\frac{5}{\Delta_S^a}\right) \le \frac{65}{\Delta_S^a} \log\left(\frac{5}{\Delta_S^a}\right) \le \frac{65}{\Delta_S} \log\left(\frac{5}{\Delta_S}\right). \tag{80}$$

Therefore, combining the instance-independent regret (77) and the instance-dependent regret (80) in forced sampling, the satisficing regret incurred in forced sampling with UCB test is bounded by

$$\min\left\{\sqrt{2}c_1\sqrt{K}\gamma_i^{-1}\log\left(\frac{1}{\gamma_i}\right)^{1/2}, \frac{65}{\Delta_S}\log\left(\frac{5}{\Delta_S}\right)\right\}. \tag{81}$$

If $r(\hat{A}_i) < S$, then for every k, round i is not terminated after pulling arm $\hat{A}_i$ for k times (in addition to the number of pulls τ_i during $\texttt{ALG}(t_i)$) occurs with probability at most

$$\begin{aligned} &\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau_i+k} - \sqrt{\frac{4\log(\tau_i+k)}{\tau_i+k}} > S\right) \\ \le&\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau_i+k} - \sqrt{\frac{4\log(\tau_i+k)}{\tau_i+k}} > r(\hat{A}_i)\right) \\ \le&\exp\left(-2\cdot 4\log(\tau_i+k)\right) \\ =&\frac{1}{(\tau_i+k)^8} \le \frac{1}{(k+1)^8}, \end{aligned}$$

where the second inequality is based on Hoeffding's inequality, and the third inequality based on $\tau_i \geq 1$. Therefore expected number of arm pulls of $\hat{A}_i$ after first pulling it τ_i times during $\texttt{ALG}(t_i)$ is bounded by

$$\begin{aligned}&\mathbb{E}\left[\text{Number of times } \hat{A}_i \text{ is pulled in round } i \text{ after the first } \tau_i \text{ times}\right]\\ =&\sum_{s=1}^{\infty}\Pr(k \geq s) = \sum_{s=1}^{\infty}\frac{1}{(s+1)^8} \leq 1.\end{aligned}$$

So, the satisficing regret in LCB test can be bounded by 1.

Summing up the two parts of satisficing regret, *i.e.*, the satisficing regret from $\texttt{ALG}(t_i)$ (76), forced sampling with UCB test (81), and LCB test, we have that the satisficing regret of round i is bounded by

$$\begin{aligned}&\min\left\{2\sqrt{2}c_1\sqrt{K}\gamma_i^{-1}\log\left(\frac{1}{\gamma_i}\right)^{1/2}+1,\ \frac{65}{\Delta_S}\log\left(\frac{5}{\Delta_S}\right)+\sum_{a\in\mathcal{A},\, r(a)<S}\frac{2c_2\log(1/\gamma_i)}{\Delta_S^a}+1\right\}\\ \leq&\min\left\{4c_1\sqrt{K}\gamma_i^{-1}\log\left(\frac{1}{\gamma_i}\right)^{1/2},\ \frac{66}{\Delta_S}\log\left(\frac{5}{\Delta_S}\right)+\frac{K}{\Delta_S}\cdot 2c_2\log\left(\frac{1}{\gamma_i}\right)\right\}.\end{aligned}$$

N.4. Proof of Proposition 10

Using similar arguments as in (18) that

$$\mathbb{E}[r(A^*)-r(\hat{A}_i)] \leq \frac{1}{t_i}\cdot c_1\sqrt{Kt_i\log(t_i)} = \sqrt{2}c_1\sqrt{K}\gamma_i\log(1/\gamma_i)^{1/2}.$$

According to the condition of this proposition, we have i satisfies the condition

$$16\sqrt{2}c_1\sqrt{K}\gamma_i\log(1/\gamma_i)^{1/2} \leq \Delta_S^*.$$

By the Markov's inequality, we have

$$\Pr(r(A^*)-r(\hat{A}_i)\geq\Delta_S^*/2) \leq \Pr\left(r(A^*)-r(\hat{A}_i)\geq 8\sqrt{2}c_1\sqrt{K}\gamma_i\log(1/\gamma_i)^{1/2}\right)\leq\frac{1}{8}. \tag{82}$$

If $r(\hat{A}_i)\geq r(A^*)-\Delta_S^*/2$, which happens with probability at least 7/8 according to the above inequality, then $r(\hat{A}_i)\geq S+\Delta_S^*/2$ holds. We first analyze the exit probability in LCB test. If i satisfies the condition stated in the proposition, we can recall Inequality (20) that

$$\sqrt{\frac{4\log(T_i+k)}{T_i+k}}\leq\frac{\Delta_S^*}{4},$$

Suppose round i is terminated for some k, then

$$\frac{\hat{r}_i^{\text{tot}}}{T_i+k}-\sqrt{\frac{4\log(T_i+k)}{T_i+k}}<S=r(A^*)-\Delta_S^*\leq r(\hat{A}_i)-\frac{\Delta_S^*}{2}\leq r(\hat{A}_i)-2\sqrt{\frac{4\log(T_i+k)}{T_i+k}},$$

or equivalently

$$\frac{\hat{r}_i^{\text{tot}}}{T_i+k} \le r(\hat{A}_i) - \sqrt{\frac{4\log(T_i+k)}{T_i+k}}.$$

By the Hoeffding's inequality we have

$$\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{T_i+k} \le r(\hat{A}_i) - \sqrt{\frac{4\log(T_i+k)}{T_i+k}}\right) \le \frac{1}{(T_i+k)^8} \le \frac{1}{(k+1)^8}.$$

Therefore conditioned on $r(\hat{A}_i) \ge r(A^*) - \Delta_S^*/2$, the probability that round i is terminated during LCB test by the end of the horizon is bounded by

$$\sum_{k=1}^{\infty} \frac{1}{(k+1)^8} \le \frac{1}{16}. \tag{83}$$

Then, we consider the probability of termination during forced sampling with UCB test given $r(\hat{A}_i) \ge S + \Delta_S^*/2$. For any $\tau \in [T_i]$, we start a new round if

$$\frac{\hat{r}_i^{\text{tot}}}{\tau} + \sqrt{\frac{4\log(\tau+1)}{\tau}} < S < r(\hat{A}_i) - \Delta_S^* < r(\hat{A}_i).$$

By the Hoeffding's inequality we have

$$\Pr\left(\frac{\hat{r}_i^{\text{tot}}}{\tau} + \sqrt{\frac{4\log(\tau+1)}{\tau}} < r(\hat{A}_i)\right) \le \frac{1}{(\tau+1)^8}.$$

Therefore conditioned on $r(\hat{A}_i) \ge r(A^*) - \Delta_S^*/2$, the probability that round i is terminated during forced sampling with UCB test by the end of the horizon is bounded by

$$\sum_{\tau=1}^{T_i} \frac{1}{(\tau+1)^8} \le \sum_{\tau=1}^{\infty} \frac{1}{(\tau+1)^8} \le \frac{1}{16}. \tag{84}$$

Combining (82), (83), and (84), we conclude that the probability that round i terminates by the end of the time horizon is upper bounded by 1/4.

N.5. Proof of Theorem 13

Because the regret incurred during forced sampling with UCB test is bounded by the one incurred without UCB test, we know the the standard regret incurred in round i is still at most

$$\frac{8c_1\sqrt{K}}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(4/\gamma_i)^\beta.$$

Since the length of the time horizon is T time steps, we have $t_i \le T$, thus the maximum number of rounds is bounded by $i_{\max} = \lceil \alpha \log_2(T) \rceil$. Therefore, the same as the regret bound in Theorem 2 (following the same chain of inequality in (13)), the regret is bounded by

$$\mathbb{E}[\texttt{Regret}] \le \sum_{i=1}^{i_{\max}} \frac{8c_1\sqrt{K}}{(1-\alpha)^\beta}\gamma_i^{-\alpha/(1-\alpha)}\log(8T)^\beta$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{i_{\max}} \frac{8c_1\sqrt{K}}{(1-\alpha)^\beta} 2^{i+1} \log(8T)^\beta \\
&\leq \frac{32c_1\sqrt{K}}{(1-\alpha)^\beta} 2^{i_{\max}} \log(8T)^\beta \\
&\leq \frac{32c_1\sqrt{K}}{(1-\alpha)^\beta} 2^{\alpha \log_2(T)+1} \log(8T)^\beta \\
&\leq \frac{64c_1\sqrt{K}}{(1-\alpha)^\beta} T^\alpha \log(8T)^\beta.
\end{aligned}
$$

Plugging into $\alpha = 1/2$ and $c_1\sqrt{K} \propto \sqrt{K}$ in the UCB algorithm, the standard regret is bounded by $\sqrt{KT} \cdot \text{polylog}(T)$.

O. Dependence in Dimension for Lipschitz Bandits

In Corollary 3, the expected constant satisficing regret of `SELECT` in the realizable case is $\tilde{O}\left(\Delta_S^{*-(d+1)}\right)$, which grows exponentially in the dimension d. In this section, we test this exponential dependence on d. We consider a class of instances of Lipschitz bandits with concave expected reward in domain $(x_1, \cdots, x_d) \in [0,1]^d$. To fix the Lipschitz coefficient, the rewards is set as $r(x_1, \cdots, x_d) = 1 - 25/(7\sqrt{d}) \cdot \sum_{i=1}^d (x_i - 0.7)^2$ so that $L = 5$ for all $d \geq 1$. We vary the dimension from 1 to 7 with a stepsize of 1. The time horizon is set to be 500000 and the satisficing level is set to be $S = 0.5$. We use the Uniform UCB as the learning oracle. The experiment is repeated for 100 times and we report the average results.

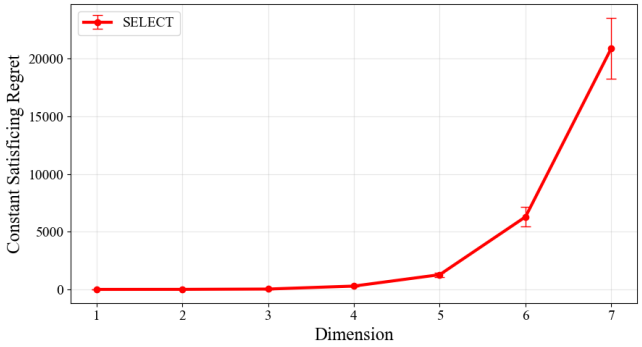


Figure 15 **Constant satisficing regret changing with dimension d in Lipschitz bandits**

The result is presented in Figure 15, where one can see that the constant satisficing regret follows an exponential increase as the dimension d gets large, which matches our theoretical result.

P. Auxiliary Results

For completeness, we restate some known results from existing literature.

Consider the following K instances of K-arm bandits. The reward of each arm follows a Bernoulli distribution, and under instance k, the mean rewards are defined as

$$r_k^{(k)} = \frac{1}{2} + \Delta,\ r_{k'}^{(k)} = \frac{1}{2},\ \forall k' \neq k. \tag{85}$$

The following lemma establishes the information impossibility of distinguishing the instances defined in (85) within $O(K/\Delta^2)$ time steps.

Lemma 1 (Lemma 2.8 of Slivkins et al. (2019))**.** *Let π be any non-anticipatory policy that commits one arm i_T after adaptively making T arm pulls. Then there exists a constant $0 < c < 1$ such that for all $T \leq cK/\Delta^2$, there exists at least $\lceil K/3 \rceil$ arms $k \in [K]$ such that $Pr^{(k)}(i_T = k) < 3/4$.*