Improving Regret Bounds for Combinatorial Semi-Bandits with Probabilistically Triggered Arms and Its Applications

Qinshi Wang
Princeton University
Princeton, NJ 08544
qinshiw@princeton.edu

Wei Chen
Microsoft Research
Beijing, China
weic@microsoft.com

Abstract

We study combinatorial multi-armed bandit with probabilistically triggered arms and semi-bandit feedback (CMAB-T). We resolve a serious issue in the prior CMAB-T studies where the regret bounds contain a possibly exponentially large factor of $1/p^*$, where p^* is the minimum positive probability that an arm is triggered by any action. We address this issue by introducing a triggering probability modulated (TPM) bounded smoothness condition into the general CMAB-T framework, and show that many applications such as influence maximization bandit and combinatorial cascading bandit satisfy this TPM condition. As a result, we completely remove the factor of $1/p^*$ from the regret bounds, achieving significantly better regret bounds for influence maximization and cascading bandits than before. Finally, we provide lower bound results showing that the factor $1/p^*$ is unavoidable for general CMAB-T problems, suggesting that the TPM condition is crucial in removing this factor.

1 Introduction

Stochastic multi-armed bandit (MAB) is a classical online learning framework modeled as a game between a player and the environment with m arms. In each round, the player selects one arm and the environment generates a reward of the arm from a distribution unknown to the player. The player observes the reward, and use it as the feedback to the player's algorithm (or policy) to select arms in future rounds. The goal of the player is to cumulate as much reward as possible over time. MAB models the classical dilemma between exploration and exploitation: whether the player should keep exploring arms in search for a better arm, or should stick to the best arm observed so far to collect rewards. The standard performance measure of the player's algorithm is the *(expected) regret*, which is the difference in expected cumulative reward between always playing the best arm in expectation and playing according to the player's algorithm.

In recent years, stochastic combinatorial multi-armed bandit (CMAB) receives many attention (e.g. [9, 7, 6, 10, 13, 15, 14, 16, 8]), because it has wide applications in wireless networking, online advertising and recommendation, viral marketing in social networks, etc. In the typical setting of CMAB, the player selects a combinatorial action to play in each round, which would trigger the play of a set of arms, and the outcomes of these triggered arms are observed as the feedback (called semi-bandit feedback). Besides the exploration and exploitation tradeoff, CMAB also needs to deal with the exponential explosion of the possible actions that makes exploring all actions infeasible.

One class of the above CMAB problems involves probabilistically triggered arms [7, 14, 16], in which actions may trigger arms probabilistically. We denote it as CMAB-T in this paper. Chen et al. [7] provide such a general model and apply it to the influence maximization bandit, which models

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

stochastic influence diffusion in social networks and sequentially selecting seed sets to maximize the cumulative influence spread over time. Kveton et al. [14, 16] study cascading bandits, in which arms are probabilistically triggered following a sequential order selected by the player as the action. However, in both studies, the regret bounds contain an undesirable factor of $1/p^*$, where p^* is the minimum positive probability that any arm can be triggered by any action,[1] and this factor could be exponentially large for both influence maximization and cascading bandits.

In this paper, we adapt the general CMAB framework of [7] in a systematic way to completely remove the factor of $1/p^*$ for a large class of CMAB-T problems including both influence maximization and combinatorial cascading bandits. The key observation is that for these problems, a harder-to-trigger arm has less impact to the expected reward and thus we do not need to observe it as often. We turn this key observation into a triggering probability modulated (TPM) bounded smoothness condition, adapted from the original bounded smoothness condition in [7]. We eliminates the $1/p^*$ factor in the regret bounds for all CMAB-T problems with the TPM condition, and show that influence maximization bandit and the conjunctive/disjunctive cascading bandits all satisfy the TPM condition. Moreover, for general CMAB-T without the TPM condition, we show a lower bound result that $1/p^*$ is unavoidable, because the hard-to-trigger arms are crucial in determining the best arm and have to be observed enough times.

Besides removing the exponential factor, our analysis is also tighter in other regret factors or constants comparing to the existing influence maximization bandit results [7, 25], combinatorial cascading bandit [16], and linear bandits without probabilistically triggered arms [15]. Both the regret analysis based on the TPM condition and the proof that influence maximization bandit satisfies the TPM condition are technically involved and nontrivial, but due to the space constraint, we have to move the complete proofs to the supplementary material. Instead we introduce the key techniques used in the main text.

Related Work. Multi-armed bandit problem is originally formated by Robbins [20], and has been extensively studied in the literature [cf. 3, 21, 4]. Our study belongs to the stochastic bandit research, while there is another line of research on adversarial bandits [2], for which we refer to a survey like [4] for further information. For stochastic MABs, an important approach is Upper Confidence Bound (UCB) approach [1], on which most CMAB studies are based upon.

As already mentioned in the introduction, stochastic CMAB has received many attention in recent years. Among the studies, we improve (a) the general framework with probabilistically triggered arms of [7], (b) the influence maximization bandit results in [7] and [25], (c) the combinatorial cascading bandit results in [16], and (d) the linear bandit results in [15]. We defer the technical comparison with these studies to Section 4.3. Other CMAB studies do not deal with probabilistically triggered arms. Among them, [9] is the first study on linear stochastic bandit, but its regret bound has since been improved by Chen et al. [7], Kveton et al. [15]. Combes et al. [8] improve the regret bound of [15] for linear bandits in a special case where arms are mutually independent. Most studies above are based on the UCB-style CUCB algorithm or its minor variant, and differ on the assumptions and regret analysis. Gopalan et al. [10] study Thompson sampling for complex actions, which is based on the Thompson sample approach [22] and can be applied to CMAB, but their regret bound has a large exponential constant term.

Influence maximization is first formulated as a discrete optimization problem by Kempe et al. [12], and has been extensively studied since (cf. [5]). Variants of influence maximization bandit have also been studied [18, 23, 24]. Lei et al. [18] use a different objective of maximizing the expected size of the union of the influenced nodes over time. Vaswani et al. [23] discuss how to transfer node level feedback to the edge level feedback, and then apply the result of [7]. Vaswani et al. [24] replace the original maximization objective of influence spread with a heuristic surrogate function, avoiding the issue of probabilistically triggered arms. But their regret is defined against a weaker benchmark relaxed by the approximation ratio of the surrogate function, and thus their theoretical result is weaker than ours.

[1] The factor of $1/f^*$ used for the combinatorial disjunctive cascading bandits in [16] is essentially $1/p^*$.

2 General Framework

In this section we present the general framework of combinatorial multi-armed bandit with probabilistically triggered arms originally proposed in [7] with a slight adaptation, and denote it as CMAB-T. We illustrate that the influence maximization bandit [7] and combinatorial cascading bandits [14, 16] are example instances of CMAB-T.

CMAB-T is described as a learning game between a learning agent (or player) and the environment. The environment consists of m random variables $X_1, \ldots, X_m$ called *base arms* (or *arms*) following a joint distribution D over $[0,1]^m$. Distribution D is picked by the environment from a class of distributions $\mathcal{D}$ before the game starts. The player knows $\mathcal{D}$ but not the actual distribution D.

The learning process proceeds in discrete rounds. In round $t \geq 1$, the player selects an action S_t from an action space $\mathcal{S}$ based on the feedback history from the previous rounds, and the environment draws from the joint distribution D an independent sample $X^{(t)} = (X_1^{(t)}, \ldots, X_m^{(t)})$. When action S_t is played on the environment outcome $X^{(t)}$, a random subset of arms $\tau_t \subseteq [m]$ are triggered, and the outcomes of $X_i^{(t)}$ for all $i \in \tau_t$ are observed as the feedback to the player. The player also obtains a nonnegative reward $R(S_t, X^{(t)}, \tau_t)$ fully determined by $S_t, X^{(t)}$, and τ_t. A learning algorithm aims at properly selecting actions S_t's over time based on the past feedback to cumulate as much reward as possible. Different from [7], we allow the action space $\mathcal{S}$ to be infinite. In the supplementary material, we discuss an example of continuous influence maximization [26] that uses continuous and infinite action space while the number of base arms is still finite.

We now describe the triggered set τ_t in more detail, which is not explicit in [7]. In general, τ_t may have additional randomness beyond the randomness of $X^{(t)}$. Let $D^{\text{trig}}(S, X)$ denote a distribution of the triggered subset of $[m]$ for a given action S and an environment outcome X. We assume that τ_t is drawn independently from $D^{\text{trig}}(S_t, X^{(t)})$. We refer D^{trig} as the *probabilistic triggering function*.

To summarize, a *CMAB-T problem instance* is a tuple $([m], \mathcal{S}, \mathcal{D}, D^{\text{trig}}, R)$, with elements already described above. These elements are known to the player, and hence establishing the problem input to the player. In contrast, the *environment instance* is the actual distribution $D \in \mathcal{D}$ picked by the environment, and is unknown to the player. The problem instance and the environment instance together form the *(learning) game instance*, in which the learning process would unfold. In this paper, we fix the environment instance D, unless we need to refer to more than one environment instances.

For each arm i, let $\mu_i = \mathbb{E}_{X \sim D}[X_i]$. Let vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ denote the expectation vector of arms. Note that vector $\boldsymbol{\mu}$ is determined by D. Same as in [7], we assume that the expected reward $\mathbb{E}[R(S, X, \tau)]$, where the expectation is taken over $X \sim D$ and $\tau \sim D^{\text{trig}}(S, X)$, is a function of action S and the expectation vector $\boldsymbol{\mu}$ of the arms. Henceforth, we denote $r_S(\boldsymbol{\mu}) \triangleq \mathbb{E}[R(S, X, \tau)]$. We remark that Chen et al. [6] relax the above assumption and consider the case where the entire distribution D, not just the mean of D, is needed to determine the expected reward. However, they need to assume that arm outcomes are mutually independent, and they do not consider probabilistically triggered arms. It might be interesting to incorporate probabilistically triggered arms into their setting, but this is out of the scope of the current paper. To allow algorithm to estimate μ_i directly from samples, we assume the outcome of an arm does not depend on whether itself is triggered, i.e. $\mathbb{E}_{X \sim D, \tau \sim D^{\text{trig}}(S,X)}[X_i \mid i \in \tau] = \mathbb{E}_{X \sim D}[X_i]$.

The performance of a learning algorithm A is measured by its *(expected) regret*, which is the difference in expected cumulative reward between always playing the best action and playing actions selected by algorithm A. Formally, let $\text{opt}_{\boldsymbol{\mu}} = \sup_{S \in \mathcal{S}} r_S(\boldsymbol{\mu})$, where $\boldsymbol{\mu} = \mathbb{E}_{X \sim D}[X]$, and we assume that $\text{opt}_{\boldsymbol{\mu}}$ is finite. Same as in [7], we assume that the learning algorithm has access to an offline (α, β)*-approximation oracle* $\mathcal{O}$, which takes $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ as input and outputs an action $S^{\mathcal{O}}$ such that $\Pr\{r_{\boldsymbol{\mu}}(S^{\mathcal{O}}) \geq \alpha \cdot \text{opt}_{\boldsymbol{\mu}}\} \geq \beta$, where α is the *approximation ratio* and β is the success probability. Under the (α, β)-approximation oracle, the benchmark cumulative reward should be the $\alpha\beta$ fraction of the optimal reward, and thus we use the following (α, β)-approximation regret:

Definition 1 ((α, β)-approximation Regret). *The T-round (α, β)-approximation regret of a learning algorithm A (using an (α, β)-approximation oracle) for a CMAB-T game instance*

$([m], \mathcal{S}, \mathcal{D}, D^{\text{trig}}, R, D)$ *with* $\boldsymbol{\mu} = \mathbb{E}_{X\sim D}[X]$ *is*

$$Reg^{A}_{\boldsymbol{\mu},\alpha,\beta}(T) = T\cdot\alpha\cdot\beta\cdot\text{opt}_{\boldsymbol{\mu}} - \mathbb{E}\left[\sum_{i=1}^{T} R(S_t^A, X^{(t)}, \tau_t)\right] = T\cdot\alpha\cdot\beta\cdot\text{opt}_{\boldsymbol{\mu}} - \mathbb{E}\left[\sum_{i=1}^{T} r_{S_t^A}(\boldsymbol{\mu})\right],$$

where S_t^A is the action A selects in round t, and the expectation is taken over the randomness of the environment outcomes $X^{(1)}, \ldots, X^{(T)}$, the triggered sets $\tau_1, \ldots, \tau_T$, as well as the possible randomness of algorithm A itself.

We remark that because probabilistically triggered arms may strongly impact the determination of the best action, but they may be hard to trigger and observe, the regret could be worse and the regret analysis is in general harder than CMAB without probabilistically triggered arms.

The above framework essentially follows [7], but we decouple actions from subsets of arms, allow action space to be infinite, and explicitly model triggered set distribution, which makes the framework more powerful in modeling certain applications (see supplementary material for more discussions).

2.1 Examples of CMAB-T: Influence Maximization and Cascading Bandits

In social influence maximization [12], we are given a weighted directed graph $G = (V, E, p)$, where V and E are sets of vertices and edges respectively, and each edge (u, v) is associated with a probability $p(u, v)$. Starting from a seed set $S \subseteq V$, influence propagates in G as follows: nodes in S are activated at time 0, and at time $t \geq 1$, a node u activated in step $t-1$ has one chance to activate its inactive out-neighbor v with an independent probability $p(u, v)$. The *influence spread* of seed set S, $\sigma(S)$, is the expected number of activated nodes after the propagation ends. The offline problem of *influence maximization* is to find at most k seed nodes in G such that the influence spread is maximized. Kempe et al. [12] provide a greedy algorithm with approximation ratio $1 - 1/e - \varepsilon$ and success probability $1 - 1/|V|$, for any $\varepsilon > 0$.

For the online influence maximization bandit [7], the edge probabilities $p(u, v)$'s are unknown and need to be learned over time through repeated influence maximization tasks: in each round t, k seed nodes S_t are selected, the influence propagation from S_t is observed, the reward is the number of nodes activated in this round, and one wants to repeat this process to cumulate as much reward as possible. Putting it into the CMAB-T framework, the set of edges E is the set of arms $[m]$, and their outcome distribution D is the joint distribution of m independent Bernoulli distributions with means $p(u, v)$ for all $(u, v) \in E$. Any seed set $S \subseteq V$ with at most k nodes is an action. The triggered arm set τ_t is the set of edges (u, v) reached by the propagation, that is, u can be reached from S_t by passing through only edges $e \in E$ with $X_e^{(t)} = 1$. In this case, the distribution $D^{\text{trig}}(S_t, X^{(t)})$ degenerates to a deterministic triggered set. The reward $R(S_t, X^{(t)}, \tau_t)$ equals to the number of nodes in V that is reached from S through only edges $e \in E$ with $X_e^{(t)} = 1$, and the expected reward is exactly the influence spread $\sigma(S_t)$. The offline oracle is a $(1 - 1/e - \varepsilon, 1/|V|)$-approximation greedy algorithm. We remark that the general triggered set distribution $D^{\text{trig}}(S_t, X^{(t)})$ (together with infinite action space) can be used to model extended versions of influence maximization, such as randomly selected seed sets in general marketing actions [12] and continuous influence maximization [26] (see supplementary material).

Now let us consider combinatorial cascading bandits [14, 16]. In this case, we have m independent Bernoulli random variables $X_1, \ldots, X_m$ as base arms. An action is to select an ordered sequence from a subset of these arms satisfying certain constraint. Playing this action means that the player reveals the outcomes of the arms one by one following the sequence order until certain stopping condition is satisfied. The feedback is the outcomes of revealed arms and the reward is a function form of these arms. In particular, in the disjunctive form the player stops when the first 1 is revealed and she gains reward of 1, or she reaches the end and gains reward 0. In the conjunctive form, the player stops when the first 0 is revealed (and receives reward 0) or she reaches the end with all 1 outcomes (and receives reward 1). Cascading bandits can be used to model online recommendation and advertising (in the disjunctive form with outcome 1 as a click) or network routing reliability (in the conjunctive form with outcome 0 as the routing edge being broken). It is straightforward to see that cascading bandits fit into the CMAB-T framework: m variables are base arms, ordered sequences are actions, and the triggered set is the prefix set of arms until the stopping condition holds.

Algorithm 1 CUCB with computation oracle.

Input: m, Oracle

For each arm i, $T_i \leftarrow 0$ {maintain the total number of times arm i is played so far}
For each arm i, $\hat{\mu}_i \leftarrow 1$ {maintain the empirical mean of X_i}
for $t = 1, 2, 3, \ldots$ **do**
  For each arm $i \in [m]$, $\rho_i \leftarrow \sqrt{\frac{3 \ln t}{2T_i}}$ {the confidence radius, $\rho_i = +\infty$ if $T_i = 0$}
  For each arm $i \in [m]$, $\bar{\mu}_i = \min\{\hat{\mu}_i + \rho_i, 1\}$ {the upper confidence bound}
  $S \leftarrow \mathsf{Oracle}(\bar{\mu}_1, \ldots, \bar{\mu}_m)$
  Play action S, which triggers a set $\tau \subseteq [m]$ of base arms with feedback $X_i^{(t)}$'s, $i \in \tau$
  For every $i \in \tau$, update T_i and $\hat{\mu}_i$: $T_i = T_i + 1$, $\hat{\mu}_i = \hat{\mu}_i + (X_i^{(t)} - \hat{\mu}_i)/T_i$
end for

3 Triggering Probability Modulated Condition

Chen et al. [7] use two conditions to guarantee the theoretical regret bounds. The first one is monotonicity, which we also use in this paper, and is restated below.

Condition 1 (Monotonicity). *We say that a CMAB-T problem instance satisfies* monotonicity*, if for any action $S \in \mathcal{S}$, for any two distributions $D, D' \in \mathcal{D}$ with expectation vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and $\boldsymbol{\mu}' = (\mu'_1, \ldots, \mu'_m)$, we have $r_S(\boldsymbol{\mu}) \le r_S(\boldsymbol{\mu}')$ if $\mu_i \le \mu'_i$ for all $i \in [m]$.*

The second condition is bounded smoothness. One key contribution of our paper is to properly strengthen the original bounded smoothness condition in [7] so that we can both get rid of the undesired $1/p^*$ term in the regret bound and guarantee that many CMAB problems still satisfy the conditions. Our important change is to use triggering probabilities to modulate the condition, and thus we call such conditions *triggering probability modulated (TPM)* conditions. The key point of TPM conditions is including the triggering probability in the condition. We use $p_i^{D,S}$ to denote the probability that action S triggers arm i when the environment instance is D. With this definition, we can also technically define p^* as $p^* = \inf_{i \in [m], S \in \mathcal{S}, p_i^{D,S} > 0} p_i^{D,S}$. In this section, we further use 1-norm based conditions instead of the infinity-norm based condition in [7], since they lead to better regret bounds for the influence maximization and cascading bandits.

Condition 2 (1-Norm TPM Bounded Smoothness). *We say that a CMAB-T problem instance satisfies 1-norm TPM bounded smoothness, if there exists $B \in \mathbb{R}^+$ (referred as the* bounded smoothness constant*) such that, for any two distributions $D, D' \in \mathcal{D}$ with expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, and any action S, we have $|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \le B \sum_{i \in [m]} p_i^{D,S} |\mu_i - \mu'_i|$.*

Note that the corresponding non-TPM version of the above condition would remove $p_i^{D,S}$ in the above condition, which is a generalization of the linear condition used in linear bandits [15]. Thus, the TPM version is clearly stronger than the non-TPM version (when the bounded smoothness constants are the same). The intuition of incorporating the triggering probability $p_i^{D,S}$ to modulate the 1-norm condition is that, when an arm i is unlikely triggered by action S (small $p_i^{D,S}$), the importance of arm i also diminishes in that a large change in μ_i only causes a small change in the expected reward $r_S(\boldsymbol{\mu})$. This property sounds natural in many applications, and it is important for bandit learning — although an arm i may be difficult to observe when playing S, it is also not important to the expected reward of S and thus does not need to be learned as accurately as others more easily triggered by S.

4 CUCB Algorithm and Regret Bound with TPM Bounded Smoothness

We use the same CUCB algorithm as in [7] (Algorithm 1). The algorithm maintains the empirical estimate $\hat{\mu}_i$ for the true mean μ_i, and feed the upper confidence bound $\bar{\mu}_i$ to the offline oracle to obtain the next action S to play. The upper confidence bound $\bar{\mu}_i$ is large if arm i is not triggered often (T_i is small), providing optimistic estimates for less observed arms. We next provide its regret bound.

Definition 2 (Gap). *Fix a distribution D and its expectation vector $\boldsymbol{\mu}$. For each action S, we define the gap $\Delta_S = \max(0, \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} - r_S(\boldsymbol{\mu}))$. For each arm i, we define*

$$\Delta^i_{\min} = \inf_{S \in \mathcal{S}: p_i^{D,S} > 0, \Delta_S > 0} \Delta_S, \qquad \Delta^i_{\max} = \sup_{S \in \mathcal{S}: p_i^{D,S} > 0, \Delta_S > 0} \Delta_S.$$

As a convention, if there is no action S such that $p_i^{D,S} > 0$ and $\Delta_S > 0$, we define $\Delta^i_{\min} = +\infty$, $\Delta^i_{\max} = 0$. We define $\Delta_{\min} = \min_{i \in [m]} \Delta^i_{\min}$, and $\Delta_{\max} = \max_{i \in [m]} \Delta^i_{\max}$.

Let $\tilde{S} = \{i \in [m] \mid p_i^{\boldsymbol{\mu},S} > 0\}$ be the set of arms that could be triggered by S. Let $K = \max_{S \in \mathcal{S}} |\tilde{S}|$. For convenience, we use $\lceil x \rceil_0$ to denote $\max\{\lceil x \rceil, 0\}$ for any real number x.

Theorem 1. *For the CUCB algorithm on a CMAB-T problem instance that satisfies monotonicity (Condition 1) and 1-norm TPM bounded smoothness (Condition 2) with bounded smoothness constant B, (1) if $\Delta_{\min} > 0$, we have distribution-dependent bound*

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \le \sum_{i \in [m]} \frac{576 B^2 K \ln T}{\Delta^i_{\min}} + \sum_{i \in [m]} \left(\left\lceil \log_2 \frac{2BK}{\Delta^i_{\min}} \right\rceil_0 + 2 \right) \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} + 4Bm; \quad (1)$$

(2) we have distribution-independent bound

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \le 12B\sqrt{mKT \ln T} + \left(\left\lceil \log_2 \frac{T}{18 \ln T} \right\rceil_0 + 2 \right) \cdot m \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} + 2Bm. \quad (2)$$

For the above theorem, we remark that the regret bounds are tight (up to a $O(\sqrt{\log T})$ factor in the case of distribution-independent bound) base on a lower bound result in [15]. More specifically, Kveton et al. [15] show that for linear bandits (a special class of CMAB-T without probabilistic triggering), the distribution-dependent regret is lower bounded by $\Omega(\frac{(m-K)K}{\Delta} \log T)$, and the distribution-independent regret is lower bounded by $\Omega(\sqrt{mKT})$ when $T \ge m/K$, for some instance where $\Delta^i_{\min} = \Delta$ for all $i \in [m]$ and $\Delta^i_{\min} < \infty$. Comparing with our regret upper bound in the above theorem, (a) for distribution-dependent bound, we have the regret upper bound $O(\frac{(m-K)K}{\Delta} \log T)$ since for that instance $B = 1$ and there are K arms with $\Delta^i_{\min} = \infty$, so tight with the lower bound in [15]; and (b) for distribution-independent bound, we have the regret upper bound $O(\sqrt{mKT \log T})$, tight to the lower bound up to a $O(\sqrt{\log T})$ factor, same as the upper bound for the linear bandits in [15]. This indicates that parameters m and K appeared in the above regret bounds are all needed. As for parameter B, we can view it simply as a scaling parameter. If we scale the reward of an instance to B times larger than before, certainly, the regret is B times larger. Looking at the distribution-dependent regret bound (Eq. (1)), $\Delta^i_{\min}$ would also be scaled by a factor of B, canceling one B factor from B^2, and $\Delta_{\max}$ is also scaled by a factor of B, and thus the regret bound in Eq. (1) is also scaled by a factor of B. In the distribution-independent regret bound (Eq. (2)), the scaling of B is more direct. Therefore, we can see that all parameters m, K, and B appearing in the above regret bounds are needed. Finally, we remark that the TPM Condition 2 can be refined such that B is replaced by arm-dependent B_i that is moved inside the summation, and B in Theorem 1 is replaced with B_i accordingly. See Appendix B.4 for details.

4.1 Novel Ideas in the Regret Analysis

Due to the space limit, the full proof of Theorem 1 is moved to the supplementary material. Here we briefly explain the novel aspects of our analysis that allow us to achieve new regret bounds and differentiate us from previous analyses such as the ones in [7] and [16, 15].

We first give an intuitive explanation on how to incorporate the TPM bounded smoothness condition to remove the factor $1/p^*$ in the regret bound. Consider a simple illustrative example of two actions S_0 and S, where S_0 has a fixed reward r_0 as a reference action, and S has a stochastic reward depending on the outcomes of its triggered base arms. Let $\tilde{S}$ be the set of arms that can be triggered by S. For $i \in \tilde{S}$, suppose i can be triggered by action S with probability p_i^S, and its true mean is μ_i and its empirical mean at the end of round t is $\hat{\mu}_{i,t}$. The analysis in [7] would need a property that, if for all $i \in \tilde{S}$ $|\hat{\mu}_{i,t} - \mu_i| \le \delta_i$ for some properly defined δ_i, then S no longer generates regrets. The analysis would conclude that arm i needs to be triggered $\Theta(\log T / \delta_i^2)$ times for the above condition

to happen. Since arm i is only triggered with probability p_i^S, it means action S may need to be played $\Theta(\log T/(p_i^S\delta_i^2))$ times. This is the essential reason why the factor $1/p^*$ appears in the regret bound.

Now with the TPM bounded smoothness, we know that the impact of $|\hat{\mu}_{i,t} - \mu_i| \le \delta_i$ to the difference in the expected reward is only $p_i^S\delta_i$, or equivalently, we could relax the requirement to $|\hat{\mu}_{i,t} - \mu_i| \le \delta_i/p_i^S$ to achieve the same effect as in the previous analysis. This translates to the result that action S would generate regret in at most $O(\log T/(p_i^S(\delta_i/p_i^S)^2)) = O(p_i^S \log T/\delta_i^2)$ rounds.

We then need to handle the case when we have multiple actions that could trigger arm i. The simple addition of $\sum_{S:p_i^S>0} p_i^S \log T/\delta_i^2$ is not feasible since we may have exponentially or even infinitely many such actions. Instead, we introduce the key idea of *triggering probability groups*, such that the above actions are divided into groups by putting their triggering probabilities p_i^S into geometrically separated bins: $(1/2, 1], (1/4, 1/2] \ldots, (2^{-j}, 2^{-j+1}], \ldots$ The actions in the same group would generate regret in at most $O(2^{-j+1}\log T/\delta_i^2)$ rounds with a similar argument, and summing up together, they could generate regret in at most $O(\sum_j 2^{-j+1}\log T/\delta_i^2) = O(\log T/\delta_i^2)$ rounds. Therefore, the factor of $1/p_i^S$ or $1/p^*$ is completely removed from the regret bound.

Next, we briefly explain our idea to achieve the improved bound over the linear bandit result in [15]. The key step is to bound regret Δ_{S_t} generated in round t. By a derivation similar to [15, 7] together with the 1-norm TPM bounded smoothness condition, we would obtain that $\Delta_{S_t} \le B\sum_{i\in\tilde{S}_t} p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i)$ with high probability. The analysis in [15] would analyze the errors $|\bar{\mu}_{i,t} - \mu_i|$ by a cascade of infinitely many sub-cases of whether there are x_j arms with errors larger than y_j with decreasing y_j, but it may still be loose. Instead we directly work on the above summation. Naive bounding the about error summation would not give a $O(\log T)$ bound because there could be too many arms with small errors. Our trick is to use a *reverse amortization*: we cumulate small errors on many sufficiently sampled arms and treat them as errors of insufficiently sample arms, such that an arm sampled $O(\log T)$ times would not contribute toward the regret. This trick tightens our analysis and leads to significantly improved constant factors.

The reverse amortization trick can be seen in Appendix B.2 Eq.(8) and the derivation that follows for the no triggered arm case, as well as in Appendix B.3, Eq. (11) in the proof of Lemma 5 for the 1-norm case.

4.2 Applications to Influence Maximization and Combinatorial Cascading Bandits

The following two lemmas show that both the cascading bandits and the influence maximization bandit satisfy the TPM condition.

Lemma 1. *For both disjunctive and conjunctive cascading bandit problem instances, 1-norm TPM bounded smoothness (Condition 2) holds with bounded smoothness constant $B = 1$.*

Lemma 2. *For the influence maximization bandit problem instances, 1-norm TPM bounded smoothness (Condition 2) holds with bounded smoothness constant $B = \tilde{C}$, where $\tilde{C}$ is the largest number of nodes any node can reach in the directed graph $G = (V, E)$.*

The proof of Lemma 1 involves a technique called *bottom-up modification*. Each action in cascading bandits can be viewed as a chain from top to bottom. When changing the means of arms below, the triggering probability of arms above is not changed. Thus, if we change $\boldsymbol{\mu}$ to $\boldsymbol{\mu}'$ backwards, the triggering probability of each arm is unaffected before its expectation is changed, and when changing the mean of an arm i, the expected reward of the action is at most changed by $p_i^{D,S}|\mu_i' - \mu_i|$.

The proof of Lemma 2 is more complex, since the bottom-up modification does not work directly on graphs with cycles. To circumvent this problem, we develop an *influence tree decomposition* technique as follows. First, we order all influence paths from the seed set S to a target v. Second, each edge is independently sampled based on its edge probability to form a random *live-edge graph*. Third, we divide the reward portion of activating v among all paths from S to v: for each live-edge graph L in which v is reachable from S, assign the probability of L to the first path from S to v in L according to the path total order. Finally, we compose all the paths from S to v into a tree with S as the root and copies of v as the leaves, so that we can do bottom-up modification on this tree and properly trace the reward changes based on the reward division we made among the paths.

4.3 Discussions and Comparisons

We now discuss the implications of Theorem 1 together with Lemmas 1 and 2 by comparing them with several existing results.

Comparison with [7] and CMAB with ∞-norm bounded smoothness conditions. Our work is a direct adaption of the study in [7]. Comparing with [7], we see that the regret bounds in Theorem 1 are not dependent on the inverse of triggering probabilities, which is the main issue in [7]. When applied to influence maximization bandit, our result is strictly stronger than that of [7] in two aspects: (a) we remove the factor of $1/p^*$ by using the TPM condition; (b) we reduce a factor of $|E|$ and $\sqrt{|E|}$ in the dominant terms of distribution-dependent and -independent bounds, respectively, due to our use of 1-norm instead of ∞-norm conditions used in Chen et al. [7]. In the supplementary material, we further provide the corresponding ∞-norm TPM bounded smoothness conditions and the regret bound results, since in general the two sets of results do not imply each other.

Comparison with [25] on influence maximization bandits. Conceptually, our work deals with the general CMAB-T framework with influence maximization and combinatorial cascading bandits as applications, while Wen et al. [25] only work on influence maximization bandit. Wen et al. [25] further study a generalization of linear transformation of edge probabilities, which is orthogonal to our current study, and could be potentially incorporated into the general CMAB-T framework. Technically, both studies eliminate the exponential factor $1/p^*$ in the regret bound. Comparing the rest terms in the regret bounds, our regret bound depends on a topology dependent term $\tilde{C}$ (Lemma 2), while their bound depends on a complicated term C_*, which is related to both topology and edge probabilities. Although in general it is hard to compare the regret bounds, for the several graph families for which Wen et al. [25] provide concrete topology-dependent regret bounds, our bounds are always better by a factor from $O(\sqrt{k})$ to $O(|V|)$, where k is the number of seeds selected in each round and V is the node set in the graph. This indicates that, in terms of characterizing the topology effect on the regret bound, our simple complexity term $\tilde{C}$ is more effective than their complicated term C_*. See Appendix D for the detailed table of comparison.

Comparison with [16] on combinatorial cascading bandits By Lemma 1, we can apply Theorem 1 to combinatorial conjunctive and disjunctive cascading bandits with bounded smoothness constant $B = 1$, achieving $O(\sum \frac{1}{\Delta^i_{\min}} K \log T)$ distribution-dependent, and $O(\sqrt{mKT \log T})$ distribution-independent regret. In contrast, besides having exactly these terms, the results in [16] have an extra factor of $1/f^*$, where $f^* = \prod_{i \in S^*} p(i)$ for conjunctive cascades, and $f^* = \prod_{i \in S^*} (1 - p(i))$ for disjunctive cascades, with S^* being the optimal solution and $p(i)$ being the probability of success for item (arm) i. For conjunctive cascades, f^* could be reasonably close to 1 in practice as argued in [16], but for disjunctive cascades, f^* could be exponentially small since items in optimal solutions typically have large $p(i)$ values. Therefore, our result completely removes the dependency on $1/f^*$ and is better than their result. Moreover, we also have much smaller constant factors owing to the new reverse amortization method described in Section 4.1.

Comparison with [15] on linear bandits. When there is no probabilistically triggered arms (i.e. $p^* = 1$), Theorem 1 would have tighter bounds since some analysis dealing with probabilistic triggering is not needed. In particular, in Eq. (1) the leading constant 624 would be reduced to 48, the $\lceil \log_2 x \rceil_0$ term is gone, and $6Bm$ becomes $2Bm$; in Eq. (2) the leading constant 50 is reduced to 14, and the other changes are the same as above (see the supplementary material). The result itself is also a new contribution, since it generalizes the linear bandit of [15] to general 1-norm conditions with matching regret bounds, while significantly reducing the leading constants (their constants are 534 and 47 for distribution-dependent and independent bounds, respectively). This improvement comes from the new reversed amortization method described in Section 4.1.

5 Lower Bound of the General CMAB-T Model

In this section, we show that there exists some CMAB-T problem instance such that the regret bound in [7] is tight, i.e. the factor $1/p^*$ in the distribution-dependent bound and $\sqrt{1/p^*}$ in the distribution-independent bound are unavoidable, where p^* is the minimum positive probability that

any base arm i is triggered by any action S. It also implies that the TPM bounded smoothness may not be applied to all CMAB-T instances.

For our purpose, we only need a simplified version of the bounded smoothness condition of [7] as below: There exists a bounded smoothness constant B such that, for every action S and every pair of mean outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, we have $|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \leq B \max_{i \in \tilde{S}} |\mu_i - \mu_i'|$, where $\tilde{S}$ is the set of arms that could possibly be triggered by S.

We prove the lower bounds using the following CMAB-T problem instance $([m], \mathcal{S}, \mathcal{D}, D^{\text{trig}}, R)$. For each base arm $i \in [m]$, we define an action S_i, with the set of actions $\mathcal{S} = \{S_1, \ldots, S_m\}$. The family of distributions $\mathcal{D}$ consists of distributions generated by every $\boldsymbol{\mu} \in [0,1]^m$ such that the arms are independent Bernoulli variables. When playing action S_i in round t, with a fixed probability p, arm i is triggered and its outcome $X_i^{(t)}$ is observed, and the reward of playing S_i is $p^{-1}X_i^{(t)}$; otherwise with probability $1-p$ no arm is triggered, no feedback is observed and the reward is 0. Following the CMAB-T framework, this means that $D^{\text{trig}}(S_i, X)$, as a distribution on the subsets of $[m]$, is either $\{i\}$ with probability p or $\varnothing$ with probability $1-p$, and the reward $R(S_i, X, \tau) = p^{-1}X_i \cdot \mathbb{I}\{\tau = \{i\}\}$. The expected reward $r_{S_i}(\boldsymbol{\mu}) = \mu_i$. So this instance satisfies the above bounded smoothness with constant $B = 1$. We denote the above instance as FTP(p), standing for fixed triggering probability instance. This instance is similar with position-based model [17] with only one position, while the feedback is different. For the FTP(p) instance, we have $p^* = p$ and $r_{S_i}(\boldsymbol{\mu}) = p \cdot p^{-1}\mu_i = \mu_i$. Then applying the result in [7], we have distributed-dependent upper bound $O(\sum_i \frac{1}{p\Delta_{\min}^i} \log T)$ and distribution-independent upper bound $O(\sqrt{p^{-1}mT \log T})$.

We first provide the distribution-independent lower bound result.

Theorem 2. *Let p be a real number with $0 < p < 1$. Then for any CMAB-T algorithm A, if $T \geq 6p^{-1}$, there exists a CMAB-T environment instance D with mean $\boldsymbol{\mu}$ such that on instance FTP(p),*

$$Reg_{\boldsymbol{\mu}}^A(T) \geq \frac{1}{170}\sqrt{\frac{mT}{p}}.$$

The proof of the above and the next theorem are all based on the results for the classical MAB problems. Comparing to the upper bound $O(\sqrt{p^{-1}mT \log T})$. obtained from [7], Theorem 2 implies that the regret upper bound of CUCB in [7] is tight up to a $O(\sqrt{\log T})$ factor. This means that the $1/p^*$ factor in the regret bound of [7] cannot be avoided in the general class of CMAB-T problems.

Next we give the distribution-dependent lower bound. For a learning algorithm, we say that it is *consistent* if, for every $\boldsymbol{\mu}$, every non-optimal arm is played $o(T^a)$ times in expectation, for any real number $a > 0$. Then we have the following distribution-dependent lower bound.

Theorem 3. *For any consistent algorithm A running on instance FTP(p) and $\mu_i < 1$ for every arm i, we have*

$$\liminf_{T \to +\infty} \frac{Reg_{\boldsymbol{\mu}}^A(T)}{\ln T} \geq \sum_{i:\mu_i < \mu^*} \frac{p^{-1}\Delta_i}{\text{kl}(\mu_i, \mu^*)},$$

where $\mu^* = \max_i \mu_i$, $\Delta_i = \mu^* - \mu_i$, and $\text{kl}(\cdot,\cdot)$ is the Kullback-Leibler divergence function.

Again we see that the distribution-dependent upper bound obtained from [7] asymptotically match the lower bound above. Finally, we remark that even if we rescale the reward from $[1, 1/p]$ back to $[0,1]$, the corresponding scaling factor B would become p, and thus we would still obtain the conclusion that the regret bounds in [7] is tight (up to a $O(\sqrt{\log T})$ factor), and thus $1/p^*$ is in general needed in those bounds.

6 Conclusion and Future Work

In this paper, we propose the TPM bounded smoothness condition, which conveys the intuition that an arm difficult to trigger is also less important in determining the optimal solution. We show that this condition is essential to guarantee low regret, and prove that important applications, such as influence maximization bandits and combinatorial cascading bandits all satisfy this condition.

There are several directions one may further pursue. One is to improve the regret bound for some specific problems. For example, for the influence maximization bandit, can we give a better algorithm or analysis to achieve a better regret bound than the one provided by the general TPM condition? Another direction is to look into other applications with probabilistically triggered arms that may not satisfy the TPM condition or need other conditions to guarantee low regret. Combining the current CMAB-T framework with the linear generalization as in [25] to achieve scalable learning result is also an interesting direction.

Acknowledgment

Wei Chen is partially supported by the National Natural Science Foundation of China (Grant No. 61433014).

References

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[3] Donald A. Berry and Bert Fristedt. *Bandit problems: Sequential Allocation of Experiments*. Chapman and Hall, 1985.

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

[5] Wei Chen, Laks V. S. Lakshmanan, and Carlos Castillo. *Information and Influence Propagation in Social Networks*. Morgan & Claypool Publishers, 2013.

[6] Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In *NIPS*, 2016.

[7] Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50):1–33, 2016. A preliminary version appeared as Chen, Wang, and Yuan, "combinatorial multi-armed bandit: General framework, results and applications", ICML'2013.

[8] Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *NIPS*, 2015.

[9] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20, 2012.

[10] Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014.

[11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[12] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 137–146, 2003.

[13] Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014.

[14] Branislav Kveton, Csaba Szepesvári, Zheng Wen, and Azin Ashkan. Cascading bandits: learning to rank in the cascade model. In *Proceedings of the 32th International Conference on Machine Learning*, 2015.

[15] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

[16] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 2015.

[17] Paul Lagrée, Claire Vernade, and Olivier Cappé. Multiple-play bandits in the position-based model. In *Advances in Neural Information Processing Systems*, pages 1597–1605, 2016.

[18] Siyu Lei, Silviu Maniu, Luyi Mo, Reynold Cheng, and Pierre Senellart. Online influence maximization. In *KDD*, 2015.

[19] Michael Mitzenmacher and Eli Upfal. *Probability and Computing*. Cambridge University Press, 2005.

[20] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.

[21] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

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

[23] Sharan Vaswani, Laks V. S. Lakshmanan, and Mark Schmidt. Influence maximization with bandits. In *NIPS Workshop on Networks in the Social and Information Sciences*, 2015.

[24] Sharan Vaswani, Branislav Kveton, Zheng Wen, Mohammad Ghavamzadeh, Laks V.S. Lakshmanan, and Mark Schmidt. Diffusion independent semi-bandit influence maximization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. to appear.

[25] Zheng Wen, Branislav Kveton, and Michal Valko. Influence maximization with semi-bandit feedback. *CoRR*, abs/1605.06593v1, 2016.

[26] Yu Yang, Xiangbo Mao, Jian Pei, and Xiaofei He. Continuous influence maximization: What discounts should we offer to social network users? In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD)*, 2016.

Supplementary Materials

A Model Discussions

A.1 Comparison with the framework of [7]

The CMAB-T framework described above essentially follows the framework of [7], but with the following noticeable differences. First, we refer to S as an abstract action from an action space $\mathcal{S}$, while in [7], S is referred to as a super arm, which is a subset of base arms $[m]$. In the case of CMAB without probabilistically triggered arms, we can simply let every super arm S be an action, and $\tau(S, X) = S$, meaning that playing super arm S deterministically triggers all and only base arms in $S \subseteq [m]$. Second, we explicitly allows action space to be infinite or even continuous space, while in [7], the action space is the subsets of base arms and thus is finite. We will see later that the infinite action space does not make essential difference in the analysis. Third, for probabilistically triggered arms, we explicitly use $\tau(S, X)$ to model them, and allows $\tau(S, X)$ to have additional randomness besides the randomness of X. In [7], probabilistic triggering is explained as further base arms being triggered based on the outcomes of previously triggered base arms, and to model certain triggering structure or additional randomness in triggering an arm, dummy base arms need to be added. However, this may require introducing a large number of dummy base arms. For example, for the cascading bandits, to specify the order of the cascade sequence, we need to add dummy base arms corresponding to every possible order of the base arms. Moreover, $\tau(S, X)$ cleanly separates the randomness known to the player from the unknown randomness from the environment outcome. For example, in the discount-based continuous influence maximization [26], $\tau(c, X)$ includes the randomness of activating the seed set from the discount vector c given by η_i's, which are known to the player. In contrast, the distribution of $X_{(u,v)}$, namely probability $p(u, v)$ on edges are unknown and need to be learned. In this case, if we use dummy base arms to model such additional triggering behavior from marketing actions to seed sets, these dummy base arms will be mixed together with edge base arms for which the learning algorithm need to learn, unless further distinction is made.

Therefore, we believe that our current adaptation CMAB-T provides a cleaner framework and is more easily to be applied to various problem instances. We remark that all the analysis and results in [7] remain unchanged with our current adaptation.

A.2 Modeling general marketing actions in influence maximization

Note that we can also use randomized $\tau(S, X)$ to model some extended versions of influence maximization. For example, general marketing actions are proposed in [12] and continuous discount actions are proposed in [26], both allowing activating seed nodes with a probability depending on the marketing intensity on the node. In particular, an action in the discount-based continuous influence maximization in [26] is a vector $c = (c_1, c_2, \ldots, c_n)$, where $c_i \in [0, 1]$ is the discount to be given to node i. Discount c_i is translated to probability $\eta_i(c_i)$ that node i is activated as a seed, where $\eta_i(\cdot)$ is a monotonically non-decreasing function with $\eta_i(0) = 0$ and $\eta_i(1) = 1$. In this case, the probabilistic triggering function $\tau(c, X)$ includes the randomness from c to seed activations based on η_i's, beyond the randomness of X. That is, even when c and X are fixed, $\tau(c, X)$ is still a random set. We further remark that in this case, the action space of all discount vectors is a continuous and infinite space, which is allowed in our adapted CMAB-T model.

B Main Regret Analysis (Proofs Related to Theorem 1)

B.1 Basics of CMAB-T problems

We utilize the following well known tail bound in our analysis.

Fact 1 (Hoeffding's Inequality [11]). *Let $X_1, \cdots, X_n$ be independent and identically distributed random variables with common support $[0, 1]$ and mean μ. Let $Y = X_1 + \cdots + X_n$. Then for all $\delta \geq 0$,*

$$\Pr\{|Y - n\mu| \geq \delta\} \leq 2e^{-2\delta^2/n}.$$

Fact 2 (Multiplicative Chernoff Bound [19]). [2] *Let $X_1, \cdots, X_n$ be Bernoulli random variables taking values from $\{0,1\}$, and $\mathbb{E}[X_t | X_1, \cdots, X_{t-1}] \geq \mu$ for every $t \leq n$. Let $Y = X_1 + \cdots + X_n$. Then for all $0 < \delta < 1$,*

$$\Pr\{Y \leq (1-\delta)n\mu\} \leq e^{-\frac{\delta^2 n\mu}{2}}.$$

We introduce the following definition to assist our analysis.

Definition 3 (Event-Filtered Regret). *For any series of events $\{\mathcal{E}_t\}_{t\geq 1}$ indexed by round number t, we define $Reg^A_{\boldsymbol{\mu},\alpha}(T, \{\mathcal{E}_t\}_{t\geq 1})$ as the regret filtered by events $\{\mathcal{E}_t\}_{t\geq 1}$, that is, regret is only counted in round t if $\mathcal{E}_t$ happens in round t. Formally,*

$$Reg^A_{\boldsymbol{\mu},\alpha}(T, \{\mathcal{E}_t\}_{t\geq 1}) = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(\mathcal{E}_t)(\alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} - r_{\boldsymbol{\mu}}(S_t^A))\right].$$

For convenience, A, α, $\boldsymbol{\mu}$ and/or T can be omitted when the context is clear, and we simply use $Reg^A_{\boldsymbol{\mu},\alpha}(T, \mathcal{E}_t)$ instead of $Reg^A_{\boldsymbol{\mu},\alpha}(T, \{\mathcal{E}_t\}_{t\geq 1})$.

The following definition describes an unlikely event that $\hat{\mu}_{i,t-1}$ is not as accurate as expected.

Definition 4. *We say that the* sampling is nice *at the beginning of round t if for every arm $i \in [m]$, $|\hat{\mu}_{i,t-1} - \mu_i| < \rho_{i,t}$, where $\rho_{i,t} = \sqrt{\frac{3\ln t}{2T_{i,t-1}}}$ in round t. Let $\mathcal{N}_t^{\mathrm{s}}$ be such event.*

Lemma 3. *For each round $t \geq 1$, $\Pr\{\neg\mathcal{N}_t^{\mathrm{s}}\} \leq 2mt^{-2}$.*

Proof. For each round $t \geq 1$, we have

$$\begin{aligned}
\Pr\{\neg\mathcal{N}_t^{\mathrm{s}}\} &= \Pr\left\{\exists i \in [m], |\hat{\mu}_{i,t-1} - \mu_i| \geq \sqrt{\frac{3\ln t}{2T_{i,t-1}}}\right\} \\
&\leq \sum_{i\in[m]} \Pr\left\{|\hat{\mu}_{i,t-1} - \mu_i| \geq \sqrt{\frac{3\ln t}{2T_{i,t-1}}}\right\}. \\
&= \sum_{i\in[m]} \sum_{k=1}^{t-1} \Pr\left\{T_{i,t-1} = k, |\hat{\mu}_{i,t-1} - \mu_i| \geq \sqrt{\frac{3\ln t}{2T_{i,t-1}}}\right\}. \qquad (3)
\end{aligned}$$

When $T_{i,t-1} = k$, $\hat{\mu}_{i,t-1}$ is the average of k i.i.d. random variables $X_i^{[1]}, \ldots, X_i^{[k]}$, where $X_i^{[j]}$ is the outcome of arm i when it is triggered for the j-th time during the execution. That is, $\hat{\mu}_{i,t-1} = \sum_{j=1}^{k} X_i^{[j]}/k$. Then we have

$$\begin{aligned}
\Pr\left\{T_{i,t-1} = k, |\hat{\mu}_{i,t-1} - \mu_i| \geq \sqrt{\frac{3\ln t}{2T_{i,t-1}}}\right\} &= \Pr\left\{T_{i,t-1} = k, \left|\sum_{j=1}^{k} X_i^{[j]}/k - \mu_i\right| \geq \sqrt{\frac{3\ln t}{2k}}\right\} \\
&\leq \Pr\left\{\left|\sum_{j=1}^{k} X_i^{[j]} - k\mu_i\right| \geq \sqrt{\frac{3k\ln t}{2}}\right\} \leq 2t^{-3}, \qquad (4)
\end{aligned}$$

where the last inequality uses the Hoeffding's Inequality (Fact 1). Combining Inequalities (3) and (4), we thus prove the lemma. □

Definition 5 (Triggering probability (TP) group). *Let i be an arm and j be a positive natural number, define the triggering probability group (of actions)*

$$\mathcal{S}_{i,j}^D = \{S \in \mathcal{S} \mid 2^{-j} < p_i^{D,S} \leq 2^{-j+1}\}.$$

Notice $\{\mathcal{S}_{i,j}^D\}_{j\geq 1}$ forms a partition of $\{S \in \mathcal{S} \mid p_i^{D,S} > 0\}$.

[2] The result in the book by [19] (Theorem 4.5 together with Exercise 4.7) only covers the case where random variables X_i's are independent. However the result can be easily generalized to our case with an almost identical proof. The only main change is to replace $\mathbb{E}\left[e^{t(\sum_{j=1}^{i-1} X_j + X_i)}\right] = \mathbb{E}\left[e^{t\sum_{j=1}^{i-1} X_j}\right]\mathbb{E}\left[e^{tX_i}\right]$ with $\mathbb{E}\left[e^{t(\sum_{j=1}^{i-1} X_j + X_i)}\right] = \mathbb{E}\left[e^{t\sum_{j=1}^{i-1} X_j}\,\mathbb{E}\left[e^{tX_i} \mid X_1, \ldots, X_{i-1}\right]\right]$.

Definition 6 (Counter). *For each TP group $\mathcal{S}_{i,j}$, we define a corresponding counter $N_{i,j}$. In a run of a learning algorithm, the counters are maintained in the following manner. All the counters are initialized to 0. In each round t, if the action S_t is chosen, then update $N_{i,j}$ to $N_{i,j}+1$ for every (i,j) that $S_t \in \mathcal{S}^D_{i,j}$. Denote $N_{i,j}$ at the end of round t with $N_{i,j,t}$. In other words, we can define the counters with the recursive equation below:*

$$N_{i,j,t} = \begin{cases} 0, & \text{if } t = 0 \\ N_{i,j,t-1}+1, & \text{if } t > 0, S_t \in \mathcal{S}^D_{i,j} \\ N_{i,j,t-1}, & \text{otherwise.} \end{cases}$$

Definition 7. *Given a series of integers $\{j^i_{\max}\}_{i\in[m]}$, we say that the* triggering is nice *at the beginning of round t (with respect to $j^i_{\max}$), if for every TP group (Definition 5) identified by arm i and $1 \le j \le j^i_{\max}$, as long as $\sqrt{\frac{6\ln t}{\frac{1}{3}N_{i,j,t-1}\cdot 2^{-j}}} \le 1$, there is $T_{i,t-1} \ge \frac{1}{3}N_{i,j,t-1}\cdot 2^{-j}$. We denote this event with $\mathcal{N}^{\mathrm{t}}_t$. It implies*

$$\rho_{i,t} = \sqrt{\frac{3\ln t}{2T_{i,t-1}}} \le \sqrt{\frac{3\ln t}{2\cdot\frac{1}{3}N_{i,j,t-1}\cdot 2^{-j}}}.$$

Lemma 4. *For a series of integers $\{j^i_{\max}\}_{i\in[m]}$, $\Pr\{\neg\mathcal{N}^{\mathrm{t}}_t\} \le \sum_{i\in[m]} j^i_{\max} t^{-2}$ for every round $t \ge 1$.*

Proof. We prove this lemma by showing $\Pr\{N_{i,j,t-1} = s, T_{i,t-1} \le \frac{1}{3}N_{i,j,t-1}\cdot 2^{-j}\} \le t^{-3}$, for $0 \le s \le t-1$ and $\sqrt{\frac{6\ln t}{s\cdot 2^{-j}}} \le 1$. Let t_k be the round that $N_{i,j}$ is increased for the k-th time, for $1 \le k \le s$. Let $Y_k = \mathbb{I}\{i \in \tau_{t_k}\}$ be a Bernoulli variable, that is, i is triggered in round t_k. When fixing the action S_{t_k}, Y_k is independent from $Y_1,\dots,Y_{k-1}$. Since $S_{t_k} \in \mathcal{S}_{i,j}$, $\mathbb{E}[Y_k \mid Y_1,\dots,Y_{k-1}] \ge 2^{-j}$. Let $Z = Y_1 + \cdots + Y_s$. By multiplicative Chernoff bound (Fact 2), we have

$$\Pr\left\{Z < \frac{1}{3}s\cdot 2^{-j}\right\} < \exp\left(-\left(\frac{2}{3}\right)^2 18\ln t/2\right) < \exp(-3\ln t) = t^{-3}.$$

By definition of T_i, there is $T_{i,t-1} \ge Z$. So $\Pr\{N_{i,j,t-1} = s, T_{i,t-1} \le \frac{1}{3}N_{i,j,t-1}\cdot 2^{-j}\} \le t^{-3}$. By taking i over $[m]$, j over $1,\dots,j^i_{\max}$, s over $0,\dots,t-1$, the lemma holds. □

B.2 The Case of No Probabilistically Triggered Arms

In this section, we state and prove a theorem for the case of no probabilistically triggered arms, i.e. $p^* = 1$, when the CMAB-T instance satisfies the 1-norm (non-TPM) bounded smoothness condition below.

Condition 3 (1-Norm Bounded Smoothness). *We say that a CMAB-T problem instance satisfies 1-norm bounded smoothness, if there exists a bounded smoothness constant $B \in \mathbb{R}^+$ such that, for any two distributions $D, D' \in \mathcal{D}$ with expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, and any action S, we have $|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \le B\sum_{i\in\tilde{S}}|\mu_i - \mu'_i|$, where $\tilde{S}$ is the set of arms that are triggered by S.*

As discussed in the main text, this theorem provides better bounds than Theorem 1 with probabilistically triggered arms. Its proof is also simpler, so the readers could choose to either get oneself familiar with the analysis with this proof first, or directly jump to the next section for the proof of Theorem 1.

Theorem 4. *For the CUCB algorithm on a CMAB (without triggering, i.e. $p^* = 1$) problem that satisfies 1-norm bounded smoothness (Condition 3) with bounded smoothness constant B,*

1. *if $\Delta_{\min} > 0$, we have distribution-dependent bound*

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \le \sum_{i\in[m]} \frac{48B^2K\ln T}{\Delta^i_{\min}} + 2Bm + \frac{\pi^2}{3}\cdot m\cdot\Delta_{\max}; \tag{5}$$

2. *we have distribution-independent bound*

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \le 14B\sqrt{KmT\ln T} + 2Bm + \frac{\pi^2}{3}\cdot m\cdot\Delta_{\max}; \tag{6}$$

Proof of Theorem 4. To unify the proofs for distribution-dependent and distribution-independent bounds, we introduce a positive real number M_i for each arm i. Let $\mathcal{F}_t$ be the event $\{r_{S_t}(\bar{\boldsymbol{\mu}}) < \alpha \cdot \mathrm{opt}(\bar{\boldsymbol{\mu}})\}$. In other words, $\mathcal{F}_t$ means the oracle fails in round t. By assumption, $\Pr\{\mathcal{F}_t\} \leq 1-\beta$. Define $M_S = \max_{i\in\tilde{S}} M_i$ for each action S, specifically, $M_S = 0$ if $\tilde{S} = \varnothing$. Define

$$\kappa_T(M,s) = \begin{cases} 2B, & \text{if } s = 0, \\ 2B\sqrt{\frac{6\ln T}{s}}, & \text{if } 1 \leq s \leq \ell_T(M), \\ 0, & \text{if } s \geq \ell_T(M)+1, \end{cases}$$

where

$$\ell_T(M) = \left\lfloor \frac{24B^2K^2\ln T}{M^2} \right\rfloor.$$

We then show that if $\{\Delta_{S_t} \geq M_{S_t}\}$, $\neg\mathcal{F}_t$ and $\mathcal{N}_t^{\mathrm{s}}$ hold, we have

$$\Delta_{S_t} \leq \sum_{i\in\tilde{S}_t} \kappa_T(M_i, T_{i,t-1}). \tag{7}$$

The right hand side of the inequality is non-negative, so it holds naturally if $\Delta_{S_t} = 0$. We only need to consider $\Delta_{S_t} > 0$. By $\mathcal{N}_t^{\mathrm{s}}$ and $\neg\mathcal{F}_t$, we have

$$r_{S_t}(\bar{\boldsymbol{\mu}}_t) \geq \alpha \cdot \mathrm{opt}(\bar{\boldsymbol{\mu}}_t) \geq \alpha \cdot \mathrm{opt}(\boldsymbol{\mu}) = r_{S_t}(\boldsymbol{\mu}) + \Delta_{S_t},$$

Then by Condition 2,

$$\Delta_{S_t} \leq r_{S_t}(\bar{\boldsymbol{\mu}}_t) - r_{S_t}(\boldsymbol{\mu}) \leq B\sum_{i\in\tilde{S}_t}(\bar{\mu}_{i,t} - \mu_i).$$

We are going to bound Δ_{S_t} by bounding $\bar{\mu}_{i,t} - \mu_i$. But before doing so, we first perform a transformation. As we have $\Delta_{S_t} \geq M_{S_t}$, so $B\sum_{i\in\tilde{S}_t}(\bar{\mu}_{i,t}-\mu_i) \geq \Delta_{S_t} \geq M_{S_t}$. We have

$$\begin{aligned}
\Delta_{S_t} &\leq B\sum_{i\in\tilde{S}_t}(\bar{\mu}_{i,t}-\mu_i) \\
&\leq -M_{S_t} + 2B\sum_{i\in\tilde{S}_t}(\bar{\mu}_{i,t}-\mu_i) \\
&= 2B\sum_{i\in\tilde{S}_t}\left[(\bar{\mu}_{i,t}-\mu_i) - \frac{M_{S_t}}{2B\left|\tilde{S}_t\right|}\right] \\
&\leq 2B\sum_{i\in\tilde{S}_t}\left[(\bar{\mu}_{i,t}-\mu_i) - \frac{M_{S_t}}{2BK}\right] \\
&\leq 2B\sum_{i\in\tilde{S}_t}\left[(\bar{\mu}_{i,t}-\mu_i) - \frac{M_i}{2BK}\right].
\end{aligned} \tag{8}$$

By $\mathcal{N}_t^{\mathrm{s}}$, we have $\bar{\mu}_{i,t} - \mu_i \leq \min\{2\rho_{i,t}, 1\}$. So

$$(\bar{\mu}_{i,t}-\mu_i) - \frac{M_i}{2BK} \leq \min\{2\rho_{i,t},1\} - \frac{M_i}{2BK} \leq \min\left\{2\sqrt{\frac{3\ln T}{2T_{i,t-1}}}, 1\right\} - \frac{M_i}{2BK}.$$

If $T_{i,t-1} \leq \ell_T(M_i)$, we have $(\bar{\mu}_{i,t}-\mu_i) - \frac{M_i}{2BK} \leq \min\left\{2\sqrt{\frac{3\ln T}{2T_{i,t-1}}}, 1\right\} \leq \frac{1}{2B}\kappa_T(M_i, T_{i,t-1})$. If $T_{i,t-1} \geq \ell_T(M_i)+1$, then $2\sqrt{\frac{3\ln T}{2T_{i,t-1}}} \leq \frac{M_i}{2BK}$, so $(\bar{\mu}_{i,t}-\mu_i) - \frac{M_i}{2BK} \leq 0 = \frac{1}{2B}\kappa_T(M_i, T_{i,t-1})$. In conclusion, we continue (8) with

$$(8) \leq \sum_{i\in\tilde{S}_t}\kappa_T(M_i, T_{i,t-1}).$$

Then in each run,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{I}(\{\Delta_{S_t} \geq M_{S_t}\} \wedge \neg\mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}}) \cdot \Delta_{S_t} &\leq \sum_{t=1}^{T} \sum_{i \in \tilde{S}_t} \kappa_T(M_i, T_{i,t-1}) \\
&= \sum_{i \in [m]} \sum_{s=0}^{T_{i,T}} \kappa_T(M_i, s) \\
&\leq \sum_{i \in [m]} \sum_{s=0}^{\ell_T(M_i)} \kappa_T(M_i, s) \\
&= 2Bm + \sum_{i \in [m]} \sum_{s=1}^{\ell_T(M_i)} 2B\sqrt{\frac{6\ln T}{s}} \\
&\leq 2Bm + \sum_{i \in [m]} \int_{s=0}^{\ell_T(M_i)} 2B\sqrt{\frac{6\ln T}{s}} \mathrm{d}s \\
&\leq 2Bm + \sum_{i \in [m]} 4B\sqrt{6\ln T \ell_T(M_i)} \\
&\leq 2Bm + \sum_{i \in [m]} 4B\sqrt{6\ln T \cdot \frac{24B^2K^2\ln T}{M_i^2}} \\
&\leq 2Bm + \sum_{i \in [m]} \frac{48B^2K\ln T}{M_i}.
\end{aligned}$$

So

$$\begin{aligned}
Reg(\{\Delta_{S_t} \geq M_{S_t}\} \wedge \neg\mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}}) &= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}(\{\Delta_{S_t} \geq M_{S_t}\} \wedge \neg\mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}}) \cdot \Delta_{S_t}\right] \\
&\leq 2Bm + \sum_{i \in [m]} \frac{48B^2K\ln T}{M_i}.
\end{aligned}$$

By Lemma 3, $\Pr\{\neg\mathcal{N}_t^{\mathrm{s}}\} \leq 2mt^{-2}$. Then, as $Reg(\mathcal{E}_t) \leq \sum_{t=1}^{T} \Pr\{\mathcal{E}_t\}\Delta_{\max}$ by definition of filtered regret,

$$\begin{aligned}
Reg(\neg\mathcal{N}_t^{\mathrm{s}}) \leq \sum_{t=1}^{T} 2mt^{-2} \cdot \Delta_{\max} \leq \frac{\pi^2}{3} m \cdot \Delta_{\max}, \\
Reg(\mathcal{F}_t) \leq (1-\beta)T \cdot \Delta_{\max}.
\end{aligned}$$

The filtered regret with null event

$$\begin{aligned}
Reg(\{\}) &\leq Reg(\neg\mathcal{N}_t^{\mathrm{s}}) + Reg(\mathcal{F}_t) + Reg(\Delta_{S_t} < M_{S_t}) + Reg(\{\Delta_{S_t} \geq M_{S_t}\} \wedge \neg\mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}}) \\
&\leq (1-\beta)T \cdot \Delta_{\max} + \frac{\pi^2}{3} m \cdot \Delta_{\max} + 2Bm + \sum_{i \in [m]} \frac{48B^2K\ln T}{M_i} + Reg(\Delta_{S_t} < M_{S_t}).
\end{aligned}$$

By definition of filtered regret, $Reg_{\boldsymbol{\mu},\alpha,\beta}(T) = Reg(T, \{\}) - (1-\beta)T \cdot \Delta_{\max}$, so

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \leq \frac{\pi^2}{3} m \cdot \Delta_{\max} + 2Bm + \sum_{i \in [m]} \frac{48B^2K\ln T}{M_i} + Reg(\Delta_{S_t} < M_{S_t}).$$

For distribution-dependent bound, take $M_i = \Delta_{\min}^i$, then $Reg(\Delta_{S_t} < M_{S_t}) = 0$ and we have

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \leq \sum_{i \in [m]} \frac{48B^2K\ln T}{M_i} + 2Bm + \frac{\pi^2}{3} \cdot \Delta_{\max}.$$

For distribution-independent bound, take $M_i = M = \sqrt{(48B^2mK\ln T)/T}$, then $Reg(\Delta_{S_t} < M_{S_t}) \le TM$ and we have

$$\begin{aligned}
Reg_{\boldsymbol{\mu},\alpha,\beta}(T) &\le \sum_{i\in[m]} \frac{48B^2K\ln T}{M_i} + 2Bm + \frac{\pi^2}{3}m\cdot\Delta_{\max} + Reg(\Delta_{S_t} < M_{S_t}) \\
&\le \frac{48B^2mK\ln T}{M} + 2Bm + \frac{\pi^2}{3}m\cdot\Delta_{\max} + TM \\
&= 2\sqrt{48B^2mKT\ln T} + \frac{\pi^2}{3}m\cdot\Delta_{\max} + 2Bm \\
&\le 14B\sqrt{mKT\ln T} + \frac{\pi^2}{3}m\cdot\Delta_{\max} + 2Bm. \qquad \square
\end{aligned}$$

B.3 Proof of Theorem 1 (1-Norm Case Regret Bound)

We first show the distribution-dependent upper bound (Eq. (1)) and the distribution-independent upper bound below, which is a weaker version of Eq. (2):

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \le 48B\sqrt{mKT\ln T} + \left(\left\lceil \log_2\sqrt{\frac{KT}{288m\ln T}}\right\rceil_0 + 2\right)\cdot m\cdot\frac{\pi^2}{6}\cdot\Delta_{\max} + 4Bm. \tag{9}$$

We show full proof of Eq. (2) later in Section B.3.1. The proof of Eq. (9) is based on the distribution-dependent bound (Eq. (1)) similar to other analysis, and thus could be more familiar to readers and easier to follow, while Eq. (2) has better constant and requires an independent proof as given in Section B.3.1.

Let $\mathcal{F}_t$ be the event $\{r_{S_t}(\bar{\boldsymbol{\mu}}) < \alpha\cdot\mathrm{opt}(\bar{\boldsymbol{\mu}})\}$. In other words, $\mathcal{F}_t$ means the oracle fails in round t. By assumption, $\Pr\{\mathcal{F}_t\} \le 1-\beta$.

To unify the proofs for distribution-dependent and distribution-independent bounds, we introduce a positive real number M_i for each arm i. Define $M_S = \max_{i\in\tilde{S}} M_i$ for each action S, specifically, $M_S = 0$ if $\tilde{S} = \varnothing$. To prove the distribution-dependent bound, we will let $M_i = \Delta^i_{\min}$. To prove the distribution-independent bound, we will let $M_i = M = \tilde{\Theta}(T^{-1/2})$ to balance bounds for $Reg(\{\Delta_{S_t} \ge M_{S_t}\}$ and $Reg(\{\Delta_{S_t} < M_{S_t}\})$. Implement definition of $\mathcal{N}^{\mathrm{t}}_t$ (Definition 7) with $j^i_{\max} = j_{\max}(M_i) = \left\lceil \log_2\frac{2BK}{M_i}\right\rceil_0$. Define

$$\kappa_{j,T}(M,s) = \begin{cases} 4\cdot 2^{-j}B, & \text{if } s = 0, \\ 2B\sqrt{\frac{72\cdot 2^{-j}\ln T}{s}}, & \text{if } 1 \le s \le \ell_{j,T}(M), \\ 0, & \text{if } s \ge \ell_{j,T}(M)+1, \end{cases}$$

where

$$\ell_{j,T}(M) = \left\lfloor \frac{288\cdot 2^{-j}B^2K^2\ln T}{M^2}\right\rfloor,$$

and the following lemma explains that κ is the contribution to regret.

Lemma 5. *In every run of the CUCB algorithm on a problem instance that satisfies 1-norm TPM bounded smoothness (Condition 2), for any vector $\{M_i\}_{i\in[m]}$ of positive real numbers and $1 \le t \le T$, if $\{\Delta_{S_t} \ge M_{S_t}\}, \neg\mathcal{F}_t, \mathcal{N}^{\mathrm{s}}_t$ and $\mathcal{N}^{\mathrm{t}}_t$ hold, we have*

$$\Delta_{S_t} \le \sum_{i\in\tilde{S}_t} \kappa_{j_i,T}(M_i, N_{i,j_i,t-1}),$$

where j_i is the index of the TP group with $S_t \in \mathcal{S}_{i,j_i}$ (See Definition 5).

Proof. The right hand side of the inequality is non-negative, so it holds naturally if $\Delta_{S_t} = 0$. We only need to consider $\Delta_{S_t} > 0$. By $\mathcal{N}^{\mathrm{s}}_t$ and $\neg\mathcal{F}_t$, we have

$$r_{S_t}(\bar{\boldsymbol{\mu}}_t) \ge \alpha\cdot\mathrm{opt}(\bar{\boldsymbol{\mu}}_t) \ge \alpha\cdot\mathrm{opt}(\boldsymbol{\mu}) = r_{S_t}(\boldsymbol{\mu}) + \Delta_{S_t},$$

Then by Condition 2,

$$\Delta_{S_t} \leq r_{S_t}(\bar{\boldsymbol{\mu}}_t) - r_{S_t}(\boldsymbol{\mu}) \leq B \sum_{i \in \tilde{S}_t} p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i)..$$ (10)

We are going to bound Δ_{S_t} by bounding $p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i)$. But before doing so, we first perform a transformation. As we have $\Delta_{S_t} \geq M_{S_t}$, so $B\sum_{i\in\tilde{S}_t} p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) \geq \Delta_{S_t} \geq M_{S_t}$. We have

$$\begin{aligned}
\Delta_{S_t} &\leq B \sum_{i \in \tilde{S}_t} p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) \\
&\leq -M_{S_t} + 2B \sum_{i \in \tilde{S}_t} p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) \\
&= 2B \sum_{i \in \tilde{S}_t} \left[p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_{S_t}}{2B\left|\tilde{S}_t\right|} \right] \\
&\leq 2B \sum_{i \in \tilde{S}_t} \left[p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_{S_t}}{2BK} \right] \\
&\leq 2B \sum_{i \in \tilde{S}_t} \left[p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_i}{2BK} \right].
\end{aligned}$$ (11)

Then we bound $p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i)$. By $\mathcal{N}_t^{\mathrm{s}}$,

$$\bar{\mu}_{i,t} - \mu_i < 2\rho_{i,t} = 2\sqrt{\frac{3\ln t}{2T_{i,t-1}}}.$$

Both $\bar{\mu}_{i,t}$ and μ_i are in $[0,1]$, so $\bar{\mu}_{i,t} - \mu_i \leq 1$. We then bound $p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i)$ in different cases.

- *Case I:* $1 \leq j_i \leq j_{\max}^i$. Then we have $p_i^{D,S_t} \leq 2 \cdot 2^{-j_i}$. If $\sqrt{\frac{6\ln t}{\frac{1}{3}N_{i,j_i,t-1}\cdot 2^{-j_i}}} \leq 1$, by $\mathcal{N}_t^{\mathrm{t}}$,

  $$\bar{\mu}_{i,t} - \mu_i \leq 2\sqrt{\frac{3\ln t}{2T_{i,t-1}}} \leq \sqrt{\frac{6\ln t}{\frac{1}{3}N_{i,j_i,t-1}\cdot 2^{-j_i}}},$$

  so

  $$\bar{\mu}_{i,t} - \mu_i \leq \min\left\{ \sqrt{\frac{6\ln t}{\frac{1}{3}N_{i,j_i,t-1}\cdot 2^{-j_i}}}, 1 \right\},$$

  and

  $$\begin{aligned}
  &p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) \\
  &\leq 2\cdot 2^{-j_i} \cdot \min\left\{ \sqrt{\frac{6\ln t}{\frac{1}{3}N_{i,j_i,t-1}\cdot 2^{-j_i}}}, 1 \right\} \\
  &= \min\left\{ \sqrt{\frac{72\cdot 2^{-j_i}\ln T}{N_{i,j_i,t-1}}}, 2\cdot 2^{-j_i} \right\}.
  \end{aligned}$$

  If $N_{i,j_i,t-1} \geq \ell_{j_i,T}(M_i) + 1$, then $\sqrt{\frac{72\cdot 2^{-j_i}\ln T}{N_{i,j_i,t-1}}} \leq \frac{M_i}{2BK}$ and $p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_i}{2BK} \leq 0$. If $N_{i,j_i,t-1} = 0$, we use the bound $p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) \leq 2\cdot 2^{-j_i}$. Otherwise, i.e. $1 \leq N_{i,j_i,t-1} \leq \ell_{j_i,T}(M_i)$, we use $p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) \leq \sqrt{\frac{72\cdot 2^{-j_i}\ln T}{N_{i,j_i,t-1}}}$. Recall the definition of $\kappa_{j,T}(M,s)$, then, for $1 \leq j_i \leq j_{\max}^i$, we have

  $$p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_i}{2BK} \leq \frac{1}{2B}\kappa_{j_i,T}(M_i, N_{i,j_i,t-1}).$$ (12)

- *Case II:* $j_i \geq j_{\max}^i + 1 = \left\lceil \log_2 \frac{2BK}{M_i} \right\rceil_0 + 1$. Then we have

$$\begin{aligned} p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) &\leq p_i^{D,S_t} \leq 2 \cdot 2^{-j_i} \\ &\leq 2 \cdot 2^{-\log_2 \frac{2BK}{M_i} - 1} = \frac{M_i}{2BK}. \end{aligned}$$

So

$$p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_i}{2BK} \leq 0 \leq \frac{1}{2B} \kappa_{j_i,T}(M_i, N_{i,j_i,t-1}). \tag{13}$$

Combining Eq. (11), (12) and (13), we conclude the proof with

$$\begin{aligned} \Delta_{S_t} &\leq 2B \sum_{i \in \tilde{S}_t} \left[p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_i}{2BK} \right] \\ &\leq \textstyle\sum_{i \in \tilde{S}_t} \kappa_{j_i,T}(M_i, N_{i,j_i,t-1}). \end{aligned}$$

□

We remark that the proof of Lemma 5, in particular the derivation leading to Eq. (11) together with the argument in the paragraph before Eq.(12), contains the reverse amortization trick we mentioned in the main text. In particular, by the derivation of Eq. (11), the contribution of every arm i to regret Δ_{S_t} is accounted as $2B\left[p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) - \frac{M_i}{2BK}\right]$. Then by the argument in the paragraph before Eq.(12), if $N_{i,j_i,t-1} \geq \ell_{j_i,T}(M_i) + 1$, meaning that i has been triggered by actions in group j_i for at least $\ell_{j_i,T}(M_i) + 1$, its error $|\bar{\mu}_{i,t} - \mu_i|$ would be small enough such that its contribution to the regret Δ_{S_t} is not positive. This trick eliminates the need of summing up small errors from many sufficiently sampled arms, leading to a tighter regret bound. The same trick can be seen in Appendix B.2, Eq.(8) and the derivation that follows for the no triggered arm case.

Lemma 6. *For the CUCB algorithm on a problem instance that satisfies TPM bounded smoothness with 1-norm (Condition 2),*

$$Reg(\{\Delta_{S_t} \geq M_{S_t}\} \wedge \neg \mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}}) \leq \sum_{i \in [m]} \frac{576B^2K \ln T}{M_i} + 4Bm.$$

Proof. We bound $Reg(\{\Delta_{S_t} \geq M_{S_t}\} \wedge \neg \mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}})$ with Lemma 5. In every run,

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{I}(\{\Delta_{S_t} \geq M_{S_t}\} \wedge \neg \mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}}) \Delta_{S_t} &\leq \sum_{t=1}^{T} \sum_{i \in \tilde{S}_t} \kappa_{j_i,T}(M_i, N_{i,j_i,t-1}) \\ &= \sum_{i \in [m]} \sum_{j=1}^{+\infty} \sum_{s=0}^{N_{i,j,T}-1} \kappa_{j,T}(M_i, s), \end{aligned} \tag{14}$$

where (14) is due to N_{i,j_i} is increased if and only if $i \in \tilde{S}_t$. For every arm i and $j \geq 1$,

$$\begin{aligned} \sum_{s=0}^{N_{i,j,T}-1} \kappa_{j,T}(M_i, s) &\leq \sum_{s=0}^{\ell_{j,T}(M_i)} \kappa_{j,T}(M_i, s) \qquad (15) \\ &= \kappa_{j,T}(M_i, 0) + \sum_{s=1}^{\ell_{j,T}(M_i)} \kappa_{j,T}(M_i, s) \\ &= \kappa_{j,T}(M_i, 0) + \sum_{s=1}^{\ell_{j,T}(M_i)} 2B\sqrt{\frac{72 \cdot 2^{-j_i} \ln T}{s}} \\ &\leq \kappa_{j,T}(M_i, 0) + 4B\sqrt{72 \cdot 2^{-j_i} \ln T}\sqrt{\ell_{j,T}(M_i)}, \qquad (16) \end{aligned}$$

where(15) is due to $\kappa_{j,T}(s) = 0$ when $s \geq \ell_{j,T}(M) + 1$, and (16) is due to the fact that, for every natural number integer n,

$$\sum_{s=1}^{n} \sqrt{\frac{1}{s}} \leq \int_{s=0}^{n} \sqrt{\frac{1}{s}}\, \mathrm{d}s = 2\sqrt{n}.$$

By definition, $\ell_{j,T}(M_i) \le \frac{288\cdot 2^{-j_i}B^2K^2\ln T}{M_i^2}$, so

$$\begin{aligned}(16) &\le \kappa_{j,T}(M,0) + 4B\sqrt{72\cdot 2^{-j_i}\ln T}\sqrt{\frac{288\cdot 2^{-j_i}B^2K^2\ln T}{M_i^2}} \\ &= 4\cdot 2^{-j}B + \frac{576\cdot 2^{-j_i}B^2K\ln T}{M_i}.\end{aligned}$$

Then we continue (14) with

$$\begin{aligned}(14) &\le \sum_{i\in[m]}\sum_{j=1}^{+\infty}\left(4\cdot 2^{-j}B + \frac{576\cdot 2^{-j_i}B^2K\ln T}{M_i}\right) \\ &= \sum_{i\in[m]}\left[\left(4B + \frac{576B^2K\ln T}{M_i}\right)\cdot\sum_{j=1}^{+\infty}2^{-j}\right] \\ &= \sum_{i\in[m]}\left(4B + \frac{576B^2K\ln T}{M_i}\right) \\ &= \sum_{i\in[m]}\frac{576B^2K\ln T}{M_i} + 4Bm.\end{aligned}$$

By taking expectation over all possible runs,

$$\begin{aligned}Reg(\{\Delta_{S_t}\ge M_{S_t}\}\wedge\neg\mathcal{F}_t\wedge\mathcal{N}_t^{\mathrm{s}}\wedge\mathcal{N}_t^{\mathrm{t}}) &= \mathbb{E}[\mathbb{I}(\{\Delta_{S_t}\ge M\}\wedge\neg\mathcal{F}_t\wedge\mathcal{N}_t^{\mathrm{s}}\wedge\mathcal{N}_t^{\mathrm{t}})\Delta_{S_t}] \\ &\le \sum_{i\in[m]}\frac{576B^2K\ln T}{M_i} + 4Bm.\end{aligned}$$

□

Proof of Theorem 1. Recall Definition 3, the definition of event-filtered regret:

$$Reg^A_{\boldsymbol{\mu}}(T,\{\mathcal{E}_t\}_{t\ge1}) = \mathbb{E}\left[\sum_{t=1}^T\mathbb{I}(\mathcal{E}_t)(\alpha\cdot\mathrm{opt}_{\boldsymbol{\mu}} - r_{S_t^A}(\boldsymbol{\mu}))\right] = T\cdot\alpha\cdot\mathrm{opt}_{\boldsymbol{\mu}} - \mathbb{E}\left[\sum_{t=1}^T\mathbb{I}(\mathcal{E}_t)(r_{S_t^A}(\boldsymbol{\mu}))\right].$$

Then for filtered regret with null event (the event that is always true), we have $Reg(\{\}) = Reg_{\boldsymbol{\mu},\alpha,\beta} + (1-\beta)T\cdot\alpha\cdot\mathrm{opt}_{\boldsymbol{\mu}}$. We divide this filtered regret into parts as

$$\begin{aligned}Reg(\{\}) \le\ & Reg(\{\Delta_{S_t} < M_{S_t}\}) + Reg(\mathcal{F}_t) + Reg(\neg\mathcal{N}_t^{\mathrm{s}}) + Reg(\neg\mathcal{N}_t^{\mathrm{t}}) \\ &+ Reg(\{\Delta_{S_t}\ge M_{S_t}\}\wedge\neg\mathcal{F}_t\wedge\mathcal{N}_t^{\mathrm{s}}\wedge\mathcal{N}_t^{\mathrm{t}}).\end{aligned} \tag{17}$$

By definition of filtered regret, $Reg(\mathcal{E}_t) \le \sum_{t=1}^T\mathbb{I}\{\mathcal{E}_t\}\Delta_{S_t} \le \sum_{t=1}^T\Pr\{\mathcal{E}_t\}\cdot\Delta_{\max}$, then

$$Reg(\mathcal{F}_t) \le \sum_{t=1}^T\Pr\{\mathcal{F}_t\}\Delta_{\max} = (1-\beta)T\cdot\Delta_{\max}, \tag{18}$$

$$Reg(\neg\mathcal{N}_t^{\mathrm{s}}) \le \sum_{t=1}^T\Pr\{\neg\mathcal{N}_t^{\mathrm{s}}\}\Delta_{\max} \le \frac{\pi^2}{3}\cdot m\cdot\Delta_{\max}, \tag{19}$$

$$Reg(\neg\mathcal{N}_t^{\mathrm{t}}) \le \sum_{t=1}^T\Pr\{\neg\mathcal{N}_t^{\mathrm{t}}\}\Delta_{\max} \le \frac{\pi^2}{6}\cdot\sum_{i\in[m]}j^i_{\max}\cdot\Delta_{\max}. \tag{20}$$

By Lemma 6,

$$Reg(\{\Delta_{S_t}\ge M_{S_t}\}\wedge\neg\mathcal{F}_t\wedge\mathcal{N}_t^{\mathrm{s}}\wedge\mathcal{N}_t^{\mathrm{t}}) \le \sum_{i\in[m]}\frac{576B^2K\ln T}{M_i} + 4Bm.$$

Take $M_i = \Delta^i_{\min}$. If $\Delta_{S_t} < M_{S_t}$, then $\Delta_{S_t} = 0$, since we have either $\tilde{S}_t = \varnothing$ or $\Delta_{S_t} < M_{S_t} \le M_i$ for some $i \in \tilde{S}_t$. So $Reg(\{\Delta_{S_t} < M_{S_t}\}) = 0$. Then we have

$$Reg(\{\}) \le (1-\beta)T \cdot \Delta_{\max} + \sum_{i\in[m]} \frac{576B^2K\ln T}{\Delta^i_{\min}} + 4Bm + \frac{\pi^2}{6} \cdot \sum_{i\in[m]} \left(j_{\max}(\Delta^i_{\min}) + 2\right) \cdot \Delta_{\max}, \tag{21}$$

where we abuse the notation of $j_{\max}(M) = \left\lceil \log_2 \frac{2BK}{M_i} \right\rceil_0$.

On the other hand, take $M_i = M = \sqrt{(576B^2mK\ln T)/T}$, then Δ_{S_t} is also M for every action S_t that $\tilde{S}_t$ is non-empty. We bound $Reg(\{\Delta_{S_t} < M_\})$ with

$$Reg(\{\Delta_{S_t} < M_{S_t}\}) = \sum_{t=1}^{T} \mathbb{I}\{\Delta_{S_t} < M_{S_t}\}\Delta_{S_t} \le \sum_{t=1}^{T} \mathbb{I}\{\Delta_{S_t} < M_{S_t}\}M \le TM.$$

So the filtered regret with null event is bounded by

$$\begin{aligned} Reg(\{\}) &\le (1-\beta)T \cdot \Delta_{\max} + \frac{576B^2mK\ln T}{M} + 4Bm + TM + \frac{\pi^2}{6} \cdot (j_{\max}(M)+2) \cdot m \cdot \Delta_{\max} \\ &= (1-\beta)T \cdot \Delta_{\max} + \frac{576B^2mK\ln T}{\sqrt{(576B^2mK\ln T)/T}} + 4Bm + T\sqrt{(576B^2mK\ln T)/T} \\ &\quad + \frac{\pi^2}{6} \cdot (j_{\max}(M)+2) \cdot m \cdot \Delta_{\max} \\ &\le (1-\beta)T \cdot \Delta_{\max} + 48B\sqrt{mKT\ln T} + 4Bm + \frac{\pi^2}{6} \cdot (j_{\max}(M)+2) \cdot m \cdot \Delta_{\max}. \end{aligned} \tag{22}$$

Since $Reg_{\boldsymbol{\mu},\alpha,\beta} = Reg(\{\}) - (1-\beta)T \cdot \alpha \cdot \mathrm{opt}_{\boldsymbol{\mu}} \le Reg(\{\}) - (1-\beta)T \cdot \Delta_{\max}$, (21) implies (1) and (22) implies (9). □

B.3.1 Further improvement on distribution-independent upper bound

We now prove the tighter distribution-independent bound (Eq. (2)) without going through distribution-dependent bound. We start with

$$\Delta_{S_t} \le B \sum_{i\in\tilde{S}_t} p_i^{D,S_t}(\bar{\mu}_{i,t} - \mu_i) \le B \sum_{i\in\tilde{S}_t} p_i^{D,S_t} \min\left\{1, 2\sqrt{\frac{3\ln T}{2T_{i,t-1}}}\right\}, \tag{10}$$

when events $\neg\mathcal{F}_t$ and $\mathcal{N}^s_t$ are true. Use $j_{\max} = \left\lceil \log_2 \frac{T}{18\ln T} \right\rceil_0$ to define $\mathcal{N}^{\mathrm{t}}_t$. When $\mathcal{N}^{\mathrm{t}}_t$, $\sqrt{\frac{3\ln T}{2T_{i,t-1}}} \le \sqrt{\frac{18\cdot 2^{-j_i}\ln T}{N_{i,j_i,t-1}}}$ if $j_i \le j_{\max}$ by definition of $\mathcal{N}^{\mathrm{t}}_t$, then $p_i^{D,S_t} \min\left\{1, 2\sqrt{\frac{3\ln T}{2T_{i,t-1}}}\right\} \le \min\left\{2^{-j_i+1}, \sqrt{\frac{72\cdot 2^{-j_i}\ln T}{N_{i,j_i,t-1}}}\right\}$ as $p_i^{D,S_t} \le 2^{-j_i+1}$. If $j_i > j_{\max}$, we still have $p_i^{D,S_t} \le 2^{-j_i+1}$. Because $N_{i,j_i,t-1} < T$, we have $2^{j_i+1} \ge \sqrt{\frac{72\cdot 2^{-j_i}\ln T}{N_{i,j_i,t-1}}}$. The conclusion is

$$p_i^{D,S_t} \min\left\{1, 2\sqrt{\frac{3\ln T}{2T_{i,t-1}}}\right\} \le \min\left\{2^{-j_i+1}, \sqrt{\frac{72\cdot 2^{-j_i}\ln T}{N_{i,j_i,t-1}}}\right\} \tag{23}$$

always holds, regardless $j \le j_{\max}$ or $j > j_{\max}$. So we define κ as following in this proof:

$$\kappa_{j,T}(s) = \min\left\{2B\cdot 2^{-j}, B\sqrt{\frac{72\cdot 2^{-j}\ln T}{s}}\right\}.$$

According to (10) and (23),

$$\begin{aligned} Reg(\neg\mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}}) &\le \sum_{t=1}^{T} \mathbb{I}(\neg\mathcal{F}_t \wedge \mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}}) \Delta_{S_t} \\ &\le \sum_{t=1}^{T} \sum_{i \in \tilde{S}_t} \kappa_{j_i,T}(N_{i,j_i,t-1}) \\ &= \sum_{i\in[m]} \sum_{j=1}^{+\infty} \sum_{s=0}^{N_{i,j,T}-1} \kappa_{j,T}(s). \end{aligned} \tag{24}$$

In each round, there are at most K of the counters $\{N_{i,j}\}_{i\in[m],j\in\mathbb{N}^+}$ are increased by 1, so $\sum_{i\in[m]}\sum_{j=1}^{+\infty} N_{i,j,T} \le KT$. To maximize the right hand side of (24) is to choose KT largest elements from the multiset $\{\kappa_{j,T}(s)\}_{i\in[m],j\in\mathbb{N}^+,s\in\mathbb{N}}$, consider the continuous version below which is more tractable than finding KT largest elements:

$$\begin{aligned} \sum_{i\in[m]} \sum_{j=1}^{+\infty} \sum_{s=0}^{N_{i,j,T}-1} \kappa_{j,T}(s) &\le \sum_{i\in[m]} \sum_{j=1}^{+\infty} \left(\kappa_{j,T}(0) + \sum_{s=1}^{\max\{0,N_{i,j,T}-1\}} \kappa_{j,T(s)} \right) \\ &\le 2Bm + \sum_{i\in[m]} \sum_{j=1}^{+\infty} \int_{s=0}^{N_{i,j,T}} \kappa_{j,T}(s)\mathrm{d}s \\ &\le 2Bm + \max_{\sum_{i,j} x_{i,j} \le KT} \left[\sum_{i\in[m]} \sum_{j=1}^{+\infty} \int_{s=0}^{x_{i,j}} B\sqrt{\frac{72\cdot 2^{-j}\ln T}{s}}\mathrm{d}s \right]. \end{aligned} \tag{25}$$

To maximize the above sum of integral, we must have $B\sqrt{\frac{72\cdot 2^{-j}\ln T}{x_{i,j}}} = B\sqrt{\frac{72\cdot 2^{-j'}\ln T}{x_{i',j'}}}$ for every $i, i' \in m$, $j, j' \in \mathbb{N}^+$. The solution is $x_{i,j} = 2^{-j}KT/m$. By taking the solution into (25), we have

$$\begin{aligned} (25) &= 2Bm + \sum_{i\in[m]} \sum_{j=1}^{+\infty} \int_{s=0}^{2^{-j}KT/m} B\sqrt{\frac{72\cdot 2^{-j}\ln T}{s}}\mathrm{d}s \\ &= 2Bm + \sum_{i\in[m]} \sum_{j=1}^{+\infty} B\sqrt{144\cdot 2^{-j}\cdot 2^{-j}KT\ln T/m} \\ &= 2Bm + 12B\sqrt{mKT\ln T}. \end{aligned} \tag{26}$$

Combining with Lemmas 3 & 4, we have

$$Reg(\{\}) \le (1-\beta)T\cdot\Delta_{\max} + 12B\sqrt{mKT\ln T} + \left(\left\lceil \log_2 \frac{T}{18\ln T} \right\rceil_0 + 2 \right) \cdot m \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} + 2Bm,$$

implying (2).

B.4 Refining Parameter B

We can refine 1-norm bounded smoothness (Condition 3) by replacing the parameter B with a separate parameter B_i for each arm i.

Condition 4 (Refined 1-Norm TPM Bounded Smoothness). *We say that a CMAB-T problem instance satisfies refined 1-norm TPM bounded smoothness, if there exists $B_i \in \mathbb{R}^+$ for every arm i (referred as the* bounded smoothness constant*) such that, for any two distributions $D, D' \in \mathcal{D}$ with expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, and any action S, we have $|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \le \sum_{i\in[m]} B_i p_i^{D,S} |\mu_i - \mu_i'|$.*

Then in Theorem 1, we may replace B with B_i in distribution-dependent bound and replace $B\sqrt{m}$ with $\sqrt{\sum_{i\in[m]} B_i^2}$ in distribution-independent bound, except that for the last constant term we replace

Bm with $\sum_{i\in[m]} B_i$. More specifically, we have (1) if $\Delta_{\min} > 0$, we have distribution-dependent bound

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \le \sum_{i\in[m]} \frac{576B_i^2 K \ln T}{\Delta^i_{\min}} + \sum_{i\in[m]} \left(\left\lceil \log_2 \frac{2B_i K}{\Delta^i_{\min}} \right\rceil_0 + 2 \right) \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} + 4 \sum_{i\in[m]} B_i; \tag{27}$$

(2) we have distribution-independent bound

$$Reg_{\boldsymbol{\mu},\alpha,\beta}(T) \le 12 \sqrt{\sum_{i\in[m]} B_i^2 KT \ln T} + \left(\left\lceil \log_2 \frac{T}{18 \ln T} \right\rceil_0 + 2 \right) \cdot m \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} + 2 \sum_{i\in[m]} B_i. \tag{28}$$

The proof of this refinement is almost straightforward replacement of B with B_i, except a few points that we want to highlight. The definition of κ and ℓ will be

$$\kappa_{i,j,T}(M,s) = \begin{cases} 4 \cdot 2^{-j} B_i, & \text{if } s = 0, \\ 2B_i \sqrt{\frac{72 \cdot 2^{-j} \ln T}{s}}, & \text{if } 1 \le s \le \ell_{i,j,T}(M), \\ 0, & \text{if } s \ge \ell_{i,j,T}(M) + 1, \end{cases}$$

where

$$\ell_{i,j,T}(M) = \left\lfloor \frac{288 \cdot 2^{-j} B_i^2 K^2 \ln T}{M^2} \right\rfloor.$$

To maximize the sum of integral in (25) (with B replaced by B_i), we need $B_i \sqrt{\frac{72 \cdot 2^{-j} \ln T}{x_{i,j}}} = B_{i'} \sqrt{\frac{72 \cdot 2^{-j'} \ln T}{x_{i',j'}}}$ for every $i, i' \in [m]$ and $j, j' \in \mathbb{N}^+$. So $x_{i,j} \propto 2^{-j} B_i^2$, and then $x_{i,j} = 2^{-j} B_i^2 KT / \sum_{i\in[m]} B_i^2$.

C Proofs for Applications of CMAB-T (Lemmas 1 and 2 in Section 4.2)

C.1 Proof of Lemma 1

Proof. Let S be an action. We regard S as a permutation of k of the arms. Without loss of generality, we may assume $S = (1, \ldots, k)$ for some $k \le K$. For convenience, we use $p_i^{\boldsymbol{\mu},S}$ instead of $p_i^{D,S}$, as arms are independent Bernoulli variables so that D can be determined by $\boldsymbol{\mu}$. For an arm $i > k$, i will not be triggered by action S, and thus $p_i^{\boldsymbol{\mu},S} = 0$. The reward also does not depend on those arms. So we may only consider the arms $1, \ldots, k$. For convenience, we only list the expectations of arms in S, so that $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ and $\boldsymbol{\mu}' = (\mu'_1, \ldots, \mu'_k)$.

Informally speaking, we can change the expectation of the arms from μ_i to μ'_i, in the reverse order from k to 1. Changing the expectation of an arm j does not affect the triggering probability of an arm i ordered in front of j, i.e. $i < j$. And when changing an arm from μ_i to μ'_i, the reward changes by at most $p_i^{\boldsymbol{\mu},S} |\mu_i - \mu'_i|$. Therefore the total difference of reward is at most $\sum_{i=1}^k p_i^{\boldsymbol{\mu},S} |\mu_i - \mu'_i|$.

Formally, for the conjunctive cascading bandit, $r_S(\boldsymbol{\mu}) = \prod_{j=1}^k \mu_j$, and $p_i^{\boldsymbol{\mu},S} = \prod_{j=1}^{i-1} \mu_j$ for $i = 1, \ldots, k$. For every $j = 0, 1, \ldots, k$, let

$$\boldsymbol{\mu}^{(j)} = (\mu_1, \ldots, \mu_j, \mu'_{j+1}, \ldots, \mu'_k),$$

specifically, $\boldsymbol{\mu}^{(k)} = \boldsymbol{\mu}$, $\boldsymbol{\mu}^{(0)} = \boldsymbol{\mu}'$. Then,

$$\begin{aligned}
\left|r_S(\boldsymbol{\mu}^{(j)}) - r_S(\boldsymbol{\mu}^{(j-1)})\right| &= \left|\prod_{i=1}^{k} \mu_i^{(j)} - \prod_{i=1}^{k} \mu_i^{(j-1)}\right| \\
&= \prod_{i,i\neq j} \mu_i^{(j)} \left|\mu_j^{(j)} - \mu_j^{(j-1)}\right| \\
&\leq \prod_{i=1}^{j-1} \mu_i^{(j)} \left|\mu_j^{(j)} - \mu_j^{(j-1)}\right| \\
&= \prod_{i=1}^{j-1} \mu_i \left|\mu_j - \mu_j'\right| \\
&= p_j^{\boldsymbol{\mu},S} \left|\mu_j - \mu_j'\right|,
\end{aligned}$$

$$\begin{aligned}
|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| &= \left|r_S(\boldsymbol{\mu}^{(k)}) - r_S(\boldsymbol{\mu}^{(0)})\right| \\
&\leq \sum_{j=1}^{k} \left|r_S(\boldsymbol{\mu}^{(j)}) - r_S(\boldsymbol{\mu}^{(j-1)})\right| \\
&\leq \sum_{j=1}^{k} p_j^{\boldsymbol{\mu},S} \left|\mu_j - \mu_j'\right|.
\end{aligned}$$

For the disjunctive case, let $\lambda_i = 1 - \mu_i$ for $i \in [m]$. Then we have $r_S(\boldsymbol{\mu}) = 1 - \prod_{j=1}^{k} \lambda_i$, and $p_i^{\boldsymbol{\mu},S} = \prod_{j=1}^{i-1} \lambda_j$. The rest analysis follows the same pattern as the conjunctive case. □

C.2 Proof of Lemma 2

C.2.1 Sufficient Condition

In influence maximization, there is a directed graph $G = (V, E)$. For convenience, we use an edge e as the index, e.g. μ_e. In this application, action S is a set of at most k nodes, so we also interpret S as a set of nodes.

Recall TPM bounded smoothness (Condition 2). The formula that we need to satisfy is

$$|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \leq B \sum_{e\in E} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu_e'|, \tag{29}$$

where $B = \max_{u\in V} |\{v \in V \mid v \text{ can be reached from } u\}|$ for influence maximization bandit, and $p_e^{\boldsymbol{\mu},S}$ stands for $p_e^{D,S}$ as D can be uniquely determined by $\boldsymbol{\mu}$.

Let $r_S^v(\boldsymbol{\mu})$ be the probability that v is activated. We claim that if for every node v and every $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ vectors, we have

$$|r_S^v(\boldsymbol{\mu}) - r_S^v(\boldsymbol{\mu}')| \leq \sum_{e\in E} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu_e'|, \tag{30}$$

Then we have Inequality (29). The reason is as follows. First, we show that Inequality (30) holds for all $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ is equivalent to $|r_S^v(\boldsymbol{\mu}) - r_S^v(\boldsymbol{\mu}')| \leq \sum_{e\in E, e \text{ can reach } v} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu_e'|$ for all $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$. In fact, the direction from the above inequality to Inequality (30) is trivial. For the reverse direction, let $\boldsymbol{\mu}''$ be an expectation vector such that for every edge e that can reach v, $\mu_e'' = \mu_e'$, and for every edge e that cannot reach v, $\mu_e'' = \mu_e$. Since the $r_S^v(\boldsymbol{\mu}')$ is only affected by edges that can reach v, we have $r_S^v(\boldsymbol{\mu}') = r_S^v(\boldsymbol{\mu}'')$. Then, we have $|r_S^v(\boldsymbol{\mu}) - r_S^v(\boldsymbol{\mu}')| = |r_S^v(\boldsymbol{\mu}) - r_S^v(\boldsymbol{\mu}'')| \leq \sum_{e\in E} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu_e''| = \sum_{e\in E, e \text{ can reach } v} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu_e'|$. Next, assuming $|r_S^v(\boldsymbol{\mu}) - r_S^v(\boldsymbol{\mu}')| \leq \sum_{e\in E, e \text{ can reach } v} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu_e'|$

holds for all $v \in V$, we have

$$\begin{aligned}
|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| &= |\sum_{v \in V} r_S^v(\boldsymbol{\mu}) - \sum_{v \in \Gamma(S)} r_S^v(\boldsymbol{\mu}')| \\
&\leq \sum_{v \in V} |r_S^v(\boldsymbol{\mu}) - r_S^v(\boldsymbol{\mu}')| \\
&\leq \sum_{v \in V} \sum_{e \in E, e \text{ can reach } v} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu'_e| \\
&= \sum_{e \in E} \sum_{v \in V, v \text{ can be reached from } e} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu'_e| \\
&\leq B \sum_{e \in E} p_e^{\boldsymbol{\mu},S} |\mu_e - \mu'_e|.
\end{aligned}$$

Thus, Inequality (29) holds.

Furthermore, we argue that it is sufficient to show that Inequality (30) holds when (1) $\boldsymbol{\mu} \leq \boldsymbol{\mu}'$, i.e. for every edge e, $\mu_e \leq \mu'_e$,; and (2) $|S| = 1$. The first condition is a straightforward conclusion from the Monotonicity condition (Condition 1). For the second condition, we may assume the seed set S consists of only one node without loss of generality. Otherwise, we may add a super seed node s° and add edges from s° to s and let $\mu_{(s^\circ,s)} = \mu'_{(s^\circ,s)} = 1$ for every node s in S.

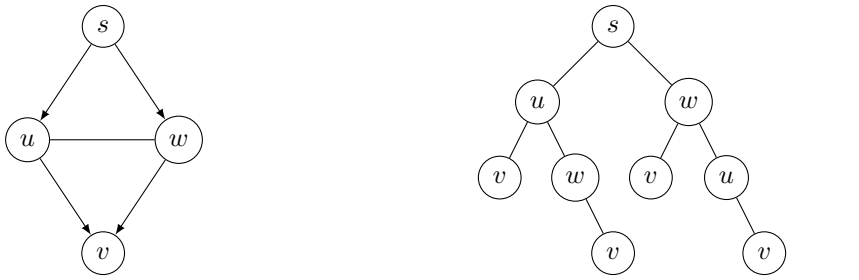
Therefore, in the rest of the proof of Lemma 2, we prove that the influence maximization bandit satisfies Inequality (30) for $\boldsymbol{\mu} \leq \boldsymbol{\mu}'$ and $|S| = 1$. Let s be the single seed node, and $S = \{s\}$.

C.2.2 Paths

In this subsection, we define an order of paths and assign the influence to the smallest path. Consider all the paths from s to v. A path L from s to v is a sequence of edges $(e_1 = (s, u_1), e_2 = (u_1, u_2), \ldots, e_{|L|} = (u_{|L|-1}, v))$. A simple path is a path that $s, v, u_1, \ldots, u_{|L|-1}$ are distinct.

We call each possible value of random vector X an outcome and denote it with vector $\boldsymbol{x} \in \{0,1\}^m$. We say an edge e is *live* (with respect to $\boldsymbol{x}$) if the corresponding component of $\boldsymbol{x}$ is 1, i.e. influence can propagate through e with the propagation under $\boldsymbol{x}$. Thus, connecting with the terminology in the influence maximization literature [12, 5], $\boldsymbol{x}$ corresponds to a *live-edge graph* in G, while X corresponds to a *random live-edge graph*. We say a path L is *live* (with respect to $\boldsymbol{x}$) if every edge of L is live. Then we have $r_S^v(\boldsymbol{\mu}) = \Pr_{\boldsymbol{x} \sim X}\{\text{there is a live path from } s \text{ to } v \text{ in } \boldsymbol{x}\}$. For each $\boldsymbol{x}$ that contains a live path from s to v, we designate a path to $\boldsymbol{x}$ as follows. We first list all the edges in an arbitrary order, and for every different edges e_1 and e_2, define $e_1 < e_2$ if e_1 appears before e_2. To compare two paths L and L', we first order the edges in L and L' in the descending order, respectively, and then compare them in the lexicographical order. In other words, to compare two paths, first compare their largest edges, if there is a tie, compare their second largest edges, and so on. If two paths continue to tie on edges and then one path ends with no more edges, then the shorter path is smaller. For every outcome $\boldsymbol{x}$ such that there is a live path from s to v, we designate the smallest live path L from s to v in $\boldsymbol{x}$ to $\boldsymbol{x}$. Then each path from s to v in the original graph G has a subset of outcome $\boldsymbol{x}$'s that are designated to L, which means all paths from s to v partition all outcomes $\boldsymbol{x}$ by which path $\boldsymbol{x}$ is designated to. Thus, let $r_{v|L}^{\boldsymbol{\mu},S} = \sum_{\boldsymbol{x} \text{ is designated to } L} \Pr[X = \boldsymbol{x}]$, namely the contribution of path L through the outcome $\boldsymbol{x}$ designated to L, and we have $r_S^v(\boldsymbol{\mu}) = \sum_{L \text{ is a path from } s \text{ to } v} r_{v|L}^{\boldsymbol{\mu},S}$. That is, we decompose $r_S^v(\boldsymbol{\mu})$ by $r_{v|L}^{\boldsymbol{\mu},S}$'s according to paths L from s to v.

Before going further, we first figure out some basic properties of the smallest live path. The smallest live path must be simple, otherwise we can remove loops to get a smaller live path. Moreover, each substring of the smallest live path in $\boldsymbol{x}$ must also be the smallest in $\boldsymbol{x}$ for its respective starting and ending nodes. For a path $L = (e_1 = (u_0, u_1), e_2 = (u_1, u_2), \ldots, e_{|L|} = (u_{|L|-1}, u_{|L|}))$, a substring is a consecutive subsequence $L_1 = (e_i, e_{i+1}, \ldots, e_j)$. If L is the smallest live path from s to v in $\boldsymbol{x}$, any substring L_1 must also be the smallest live path from u to w in $\boldsymbol{x}$, where u and w are the start and the end of L_1, respectively. Otherwise, if L_2 is a live path from u to w that smaller than L_1, then we can replace L_1 with L_2 in L to get a smaller live path.



(a) A sample network (b) Search tree with Node marked in each node

Figure 1: A sample network and its search tree

C.2.3 Bypass

In this subsection, we define bypass, which is a tool for calculating the probability that a path is *not* the smallest. For a path $L = (e_1 = (u_0, u_1), e_2 = (u_1, u_2), \ldots, e_{|L|} = (u_{|L|-1}, u_{|L|}))$, a bypass is a path from u_i to u_j that

(1) shares no edges with L;
(2) is smaller than the substring of L between u_i and u_j.

A bypass is live (with respect to $\boldsymbol{x}$) is defined in the same way as a path being live. For a live path L in $\boldsymbol{x}$ from some node u_0 to some other node $u_{|L|}$, if there is a live bypass of L, then L cannot be the smallest live path from u_0 to $u_{|L|}$. The reverse also holds: if a live path L has no live bypasses, then L is the smallest live path from u_0 to $u_{|L|}$. To prove the reverse direction, assume that there is a live path L' from u_0 to $u_{|L|}$ smaller than L. Let e_i be the largest edge in L that is not in L'. Because $L' < L$, such e_i must exist, and moreover e_i must be larger than all edges in L' but not L. By breaking L at e_i, we divide the nodes covered by L into two parts, the start part and the end part. Let w be the first node in L' that is in the end part of L. Such node w must exist because the end node $u_{|L|}$ is in the end part of L. Let u be the last node in L' that appears before w in L' and is in the start part of L. Such node u must exist because the starting node u_0 is in the start part. Then the substring of L' between u and w must share no edges with L. Otherwise, if the substring of L' between u and w shares one edge (u_j, u_{j+1}) with L, (u_j, u_{j+1}) cannot be e_i, so u cannot be u_j and w cannot be u_{j+1}. Then, (a) if u_{j+1} is in the end part of L, then u_{j+1} appearing before w in L' contradicts to w's definition; and (b) if u_{j+1} is in the start part of L, u_{j+1} appearing after u and before w in L' contradicts to the definition of u. Therefore, the substring of L' between u and w shares no edges with L. Then since e_i is larger than any edge in L' and not in L, the substring of L' between u and w is indeed a bypass of L.

For a path $L = (e_1 = (u_0, u_1), e_2 = (u_1, u_2), \ldots, e_{|L|} = (u_{|L|-1}, u_{|L|}))$, let $p_L^{\boldsymbol{\mu},S}$ be the probability that L is the smallest live path from its start to its end. Note that if L is a path from s to v, then we have $p_L^{\boldsymbol{\mu},S} = r_{v|L}^{\boldsymbol{\mu},S}$. With bypass, we have $p_L^{\boldsymbol{\mu},S} = p_{1,L}^{\boldsymbol{\mu},S} p_{2,L}^{\boldsymbol{\mu},S}$, where $p_{1,L}^{\boldsymbol{\mu},S}$ is the probability that L is live and $p_{2,L}^{\boldsymbol{\mu},S}$ is the probability that there is no live bypasses of L. It is clear that $p_{1,L}^{\boldsymbol{\mu},S} = \prod_{i=1}^{|L|} \mu_{e_i}$, and $p_{2,L}^{\boldsymbol{\mu},S}$ is the probability that some subset of edges in $E \setminus L$ forming a live bypass of L does not occur. These two events are independent, since they are about two disjoint subsets of E.

C.2.4 Bottom-up modification

We now describe the search tree formed from all simple paths from s to v. We use y, z to denote nodes in this tree. Each node y is corresponding to a prefix of a path from s to v, which is also a path denoted by $\mathrm{Path}(y)$. Denote the end node of $\mathrm{Path}(y)$ with $\mathrm{Node}(y)$. Denote the last edge of $\mathrm{Path}(y)$ with $\mathrm{Edge}(y)$. Denote the root of the tree with root. $\mathrm{Path}(\mathrm{root})$ is the empty path $\varnothing$. Specifically, $\mathrm{Node}(\mathrm{root}) = s$, as s is the start node of every path in our consideration. $\mathrm{Edge}(\mathrm{root})$ is undefined. For every non-root node y in the tree, its parent is the node z such that $\mathrm{Path}(z)$ is the $(|\mathrm{Path}(y)| - 1)$-prefix of $\mathrm{Path}(y)$. Figure 1 shows a sample of this tree structure.

For a node y in the tree, we simplify the notation $p^{\boldsymbol{\mu},S}_{\mathrm{Path}(y)}$ to $p^{\boldsymbol{\mu},S}_{y}$. Similarly, for a leaf node y in the tree, we simplify the notation $r^{\boldsymbol{\mu},S}_{v|\mathrm{Path}(y)}$ to $r^{\boldsymbol{\mu},S}_{v|y}$. Then we have $r^{v}_{S}(\boldsymbol{\mu}) = \sum_{y \text{ is leaf}} r^{\boldsymbol{\mu},S}_{v|y} = \sum_{y \text{ is a leaf}} p^{\boldsymbol{\mu},S}_{y}$.

We want to show that for all $\boldsymbol{\mu} \le \boldsymbol{\mu}'$, we have

$$r^{v}_{S}(\boldsymbol{\mu}') - r^{v}_{S}(\boldsymbol{\mu}) = \sum_{y \text{ is a leaf}} \left(p^{\boldsymbol{\mu}',S}_{y} - p^{\boldsymbol{\mu},S}_{y} \right) \le p^{\boldsymbol{\mu},S}_{e} \sum_{e \in E} (\mu'_e - \mu_e), \tag{31}$$

which is the same as Inequality (30) that we want to show.

Let $\boldsymbol{\mu}^{(y)}$ be the vector that

$$\mu^{(y)}_e = \begin{cases} \mu_e, & \text{if } e \in \mathrm{Path}(y), \\ \mu'_e, & \text{if } e \notin \mathrm{Path}(y). \end{cases}$$

Thus we have $p^{\boldsymbol{\mu}^{(y)},S}_{y} = p^{\boldsymbol{\mu},S}_{1,y} p^{\boldsymbol{\mu}',S}_{2,y}$. Since for all edges $e \notin \mathrm{Path}(y)$, $\mu_e \le \mu'_e$, the probability that there is no live bypasses of $\mathrm{Path}(y)$ is higher under $\boldsymbol{\mu}$ than under $\boldsymbol{\mu}'$, that is, $p^{\boldsymbol{\mu}',S}_{2,y} \le p^{\boldsymbol{\mu},S}_{2,y}$. Therefore, $p^{\boldsymbol{\mu}^{(y)},S}_{y} \le p^{\boldsymbol{\mu},S}_{y}$, which means that, to prove Inequality (31), it is enough to prove

$$\sum_{y \text{ is a leaf}} \left(p^{\boldsymbol{\mu}',S}_{y} - p^{\boldsymbol{\mu}^{(y)},S}_{y} \right) \le p^{\boldsymbol{\mu},S}_{e} \sum_{e \in E} (\mu'_e - \mu_e). \tag{32}$$

We now consider the bottom-up modification of the expectations in $\mathrm{Path}(y)$.

$$p^{\boldsymbol{\mu}',S}_{y} - p^{\boldsymbol{\mu}^{(y)},S}_{y} = \sum_{i=1}^{|\mathrm{Path}(y)|} \left(p^{\boldsymbol{\mu}^{(z_{i-1})},S}_{y} - p^{\boldsymbol{\mu}^{(z_i)},S}_{y} \right), \tag{33}$$

where z_i is the ancestor of y at depth i. (Root has depth 0.) By switching summations and regrouping the summands $\left(p^{\boldsymbol{\mu}^{(z_{i-1})},S}_{y} - p^{\boldsymbol{\mu}^{(z_i)},S}_{y} \right)$ under z_i, we have

$$\sum_{y \text{ is a leaf}} \left(p^{\boldsymbol{\mu}',S}_{y} - p^{\boldsymbol{\mu}^{(y)},S}_{y} \right) = \sum_{y \text{ is a non-root node}} \sum_{z \text{ is a leaf under } y} \left(p^{\boldsymbol{\mu}^{(\mathrm{Parent}(y))},S}_{z} - p^{\boldsymbol{\mu}^{(y)},S}_{z} \right). \tag{34}$$

We generalize the definition of $r^{\boldsymbol{\mu},S}_{v|y}$ to non-leaf nodes y by

$$r^{\boldsymbol{\mu},S}_{v|y} = \sum_{z \text{ is a leaf under } y} p^{\boldsymbol{\mu},S}_{z}.$$

It is clear that this definition coincides the old one when y is a leaf. Now

$$(34) = \sum_{y \text{ is a non-root node}} \left(r^{\boldsymbol{\mu}^{(\mathrm{Parent}(y))},S}_{v|y} - r^{\boldsymbol{\mu}^{(y)},S}_{v|y} \right). \tag{35}$$

$$r^{\boldsymbol{\mu},S}_{v|y} = \sum_{z \text{ is a leaf under } y} p^{\boldsymbol{\mu},S}_{z} = \sum_{z \text{ is a leaf under } y} p^{\boldsymbol{\mu},S}_{1,z} p^{\boldsymbol{\mu},S}_{2,z} = p^{\boldsymbol{\mu},S}_{1,y} \sum_{z \text{ is a leaf under } y} \frac{p^{\boldsymbol{\mu},S}_{1,z}}{p^{\boldsymbol{\mu},S}_{1,y}} p^{\boldsymbol{\mu},S}_{2,z}.$$

$\frac{p^{\boldsymbol{\mu},S}_{1,z}}{p^{\boldsymbol{\mu},S}_{1,y}} p^{\boldsymbol{\mu},S}_{2,z}$ does not depend on μ_e for every $e \in \mathrm{Path}(y)$. So

$$\begin{aligned} r^{\boldsymbol{\mu}^{(\mathrm{Parent}(y))},S}_{v|y} - r^{\boldsymbol{\mu}^{(y)},S}_{v|y} &= \left(p^{\boldsymbol{\mu}^{(\mathrm{Parent}(y))},S}_{1,y} - p^{\boldsymbol{\mu}^{(y)},S}_{1,y} \right) \sum_{z \text{ is a leaf under } y} \frac{p^{\boldsymbol{\mu}',S}_{1,z}}{p^{\boldsymbol{\mu}',S}_{1,y}} p^{\boldsymbol{\mu}',S}_{2,z} \\ &= \left(\mu'_{\mathrm{Edge}(y)} - \mu_{\mathrm{Edge}(y)} \right) p^{\boldsymbol{\mu},S}_{1,\mathrm{Parent}(y)} \sum_{z \text{ is a leaf under } y} \frac{p^{\boldsymbol{\mu}',S}_{1,z}}{p^{\boldsymbol{\mu}',S}_{1,y}} p^{\boldsymbol{\mu}',S}_{2,z}. \end{aligned} \tag{36}$$

Topology	Bound in [25]	Our bound
bar graphs	$\tilde{O}\left(\lvert V\rvert\sqrt{kT}\right)$	$\tilde{O}\left(\sqrt{k\lvert V\rvert T}\right)$
star graphs	$\tilde{O}\left(\lvert V\rvert^2\sqrt{kT}\right)$	$\tilde{O}\left(\lvert V\rvert^2\sqrt{T}\right)$
ray graphs	$\tilde{O}\left(\lvert V\rvert^{\frac{9}{4}}\sqrt{kT}\right)$	$\tilde{O}\left(\lvert V\rvert^2\sqrt{T}\right)$
tree graphs	$\tilde{O}\left(\lvert V\rvert^{\frac{5}{2}}\sqrt{T}\right)$	$\tilde{O}\left(\lvert V\rvert^2\sqrt{T}\right)$
grid graphs	$\tilde{O}\left(\lvert V\rvert^{\frac{5}{2}}\sqrt{T}\right)$	$\tilde{O}\left(\lvert V\rvert^2\sqrt{T}\right)$
complete graphs	$\tilde{O}\left(\lvert V\rvert^4\sqrt{T}\right)$	$\tilde{O}\left(\lvert V\rvert^3\sqrt{T}\right)$

Table 1: Regret bound comparison with [25].

For each leaf z under y, the event that $\mathrm{Path}(z)$ is the smallest live path from s to v is exclusive from each other. And that event is included in that $\mathrm{Path}(y)$ is the smallest live path from s to $\mathrm{Node}(y)$. So

$$\sum_{z \text{ is a leaf under } y} p_{1,z}^{\boldsymbol{\mu}',S} p_{2,z}^{\boldsymbol{\mu}',S} \le p_{1,y}^{\boldsymbol{\mu}',S} p_{2,y}^{\boldsymbol{\mu}',S},$$

and thus

$$\sum_{z \text{ is a leaf under } y} \frac{p_{1,z}^{\boldsymbol{\mu}',S}}{p_{1,y}^{\boldsymbol{\mu}',S}} p_{2,z}^{\boldsymbol{\mu}',S} \le p_{2,y}^{\boldsymbol{\mu}',S} \le p_{2,y}^{\boldsymbol{\mu},S}.$$

So

$$(36) \le \left(\mu'_{\mathrm{Edge}(y)} - \mu_{\mathrm{Edge}(y)}\right) p_{1,\mathrm{Parent}(y)}^{\boldsymbol{\mu},S} p_{2,y}^{\boldsymbol{\mu},S}.$$

Then

$$(35) \le \sum_{y \text{ is a non-root node}} \left(\mu'_{\mathrm{Edge}(y)} - \mu_{\mathrm{Edge}(y)}\right) p_{1,\mathrm{Parent}(y)}^{\boldsymbol{\mu},S} p_{2,y}^{\boldsymbol{\mu},S} = \sum_{e \in E} (\mu'_e - \mu_e) \sum_{\mathrm{Edge}(y)=e} p_{1,\mathrm{Parent}(y)}^{\boldsymbol{\mu},S} p_{2,y}^{\boldsymbol{\mu},S}. \tag{37}$$

We then show

$$\sum_{\mathrm{Edge}(y)=e} p_{1,\mathrm{Parent}(y)}^{\boldsymbol{\mu},S} p_{2,y}^{\boldsymbol{\mu},S} \le p_e^{\boldsymbol{\mu},S}, \tag{38}$$

for every edge e. If e is a directed edge from u to w , $p_e^{\boldsymbol{\mu},S} \ge \sum_{\mathrm{Edge}(y)=e} p_{\mathrm{Parent}(y)}^{\boldsymbol{\mu},S}$, since $p_{\mathrm{Parent}(y)}^{\boldsymbol{\mu},S}$ is the probability that the path $\mathrm{Path}(\mathrm{Parent}(y))$ is the smallest live path from s to $\mathrm{Node}(\mathrm{Parent}(y)) = u$, and thus such events are mutually exclusive for different y with $\mathrm{Edge}(y) = e$. Then $p_e^{\boldsymbol{\mu},S} \ge \sum_{\mathrm{Edge}(y)=e} p_{1,\mathrm{Parent}(y)}^{\boldsymbol{\mu},S} p_{2,y}^{\boldsymbol{\mu},S}$ as $p_{2,\mathrm{Parent}(y)}^{\boldsymbol{\mu},S} \ge p_{2,y}^{\boldsymbol{\mu},S}$. Thus we have (38).

Combining Inequalities (37) and (38), we prove the key Inequality (32), which in turn shows that the influence maximization bandit satisfies the TPM bounded smoothness condition with $B = \max_{u \in V} |\{v \in V \mid v \text{ can be reached from } u\}|$.

D Detailed Comparison with [25] on the Regret Bounds for Influence Maximization Bandits

Let $G = (V, E)$ be the social graph we consider. By Lemma 2, our Theorem 1 can be applied to the influence maximization bandit with $B = \tilde{C} \le |V|$, which gives concrete $O(\log T)$ distribution-dependent and $O(\sqrt{T \log T})$ distribution-independent bounds for the influence maximization bandit. Wen et al. [25] also study the influence maximization bandit and eliminate the exponential factor $1/p^*$. They use a complexity term C_* to characterize their regret bound, where C_* has complicated relationship with network topology and edge probabilities. Wen et al. [25] list several families of graphs with concrete regret bounds, ignoring the effect of edge probabilities on their complexity term C_*. Our regret bounds with complexity term $\tilde{C}$ can also be applied to these graph families, and Table 1 list the comparison results between our regret bounds and their regret bounds. The

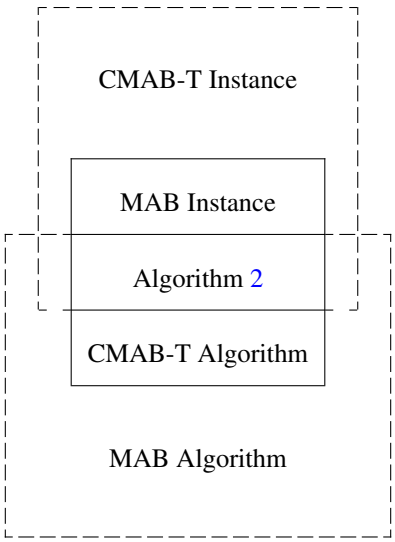


Figure 2: Reduction Structure

comparison shows that our regret bounds are always better than their bounds, with an improvement factor from $O(\sqrt{k})$ to $O(|V|)$, where V is the set of nodes in the graph, and k is the number of seeds to be selected in each round. This indicates that, in terms of characterizing the topology effect on the regret bound, our simple complexity term $\tilde{C}$ is more effective than their complicated term C_*.

E Lower Bound Proofs (for Section 5)

E.1 Proof of Theorem 2

Algorithm 2 Reduce MAB to CMAB-T

Input: m, T_{CMAB}, p {m is the number of arms, T_{CMAB} is the number of rounds in CMAB, and p is triggering probability.}

for $t = 1, \ldots, T_{\mathrm{CMAB}}$ **do**
  sample γ_t i.i.d. from Bernoulli distribution B_p
end for
$\mathcal{H} \leftarrow \varnothing$; $t_{\mathrm{MAB}} \leftarrow 0$
for $t = 1, \ldots, T_{\mathrm{CMAB}}$ **do**
  $S_{i_t} \leftarrow$ CMAB-Oracle$(\mathcal{H})$ {Oracle decides the CMAB-T action based on the execution history}
  if $\gamma_t = 1$ **then**
    $t_{\mathrm{MAB}} \leftarrow t_{\mathrm{MAB}} + 1$
    In MAB, play arm i_t in round t_{MAB}, obtain feedback $\tilde{X}_{i_t}^{(t_{\mathrm{MAB}})}$
    In CMAB-T, i_t is triggered with feedback $X_{i_t}^{(t)} = \tilde{X}_{i_t}^{(t_{\mathrm{MAB}})}$, and set reward as $p^{-1} X_{i_t}^{(t)}$
    $\mathcal{H} \leftarrow$ Append$(\mathcal{H}, (S_{i_t}, \{i_t\}, X_{i_t}^{(t)})$ {$\{i_t\}$ is the set of triggered arms}
  else
    {$\gamma_t = 0$, and MAB is not played in this case}
    In CMAB-T, no arm is triggered, and the reward is 0
    $\mathcal{H} \leftarrow$ Append$(\mathcal{H}, (S_{i_t}, \varnothing, -))$ {triggering set is empty, so no feedback}
  end if
end for {In the end, $T_{\mathrm{MAB}} = t_{\mathrm{MAB}}$}

We prove the theorem by reducing classical MAB to this CMAB-T game instance by Algorithm 2. For convenience, we define Bernoulli random variable $\gamma_t = \mathbb{I}\{\tau_t(S_{i_t}, X^{(t)}) = \{i_t\}\}$, where S_{i_t} is the action played in round t, and thus γ_t is an indicator representing whether a base arm is triggered

in round t. Moreover, to distinguish the environment outcome in MAB and CMAB-T in the reduction, we use $\tilde{X}^{(t_{\text{MAB}})}$ to denote the environment outcome in round t_{MAB} of MAB, and $X^{(t)}$ to denote the environment outcome in round t of CMAB-T.

Figure 2 shows the structure of reduction. Algorithm 2 adapts the CMAB-T algorithm to an MAB algorithm. Conversely, it also adapts the MAB instance to the corresponding CMAB-T instance. Thus when Algorithm 2 runs, we have one MAB instance and one CMAB-T instance running simultaneously. Let T_{CMAB} be the total number of rounds in the CMAB-T instance and T_{MAB} be the total number of rounds in the MAB instance. For convenience, we use t to refer to the index of rounds in CMAB-T, while t_{MAB} is the index of rounds in MAB. In Algorithm 2, we fix T_{CMAB} and thus T_{MAB} is a random variable. We have $T_{\text{MAB}} = \sum_{t=1}^{T_{\text{CMAB}}} \gamma_t$. So $\mathbb{E}[T_{\text{MAB}}] = pT_{\text{CMAB}}$ and we have following lemma about the distribution of T_{MAB}.

Lemma 7. *If $pT_{\text{CMAB}} \geq 6$, then $\Pr\left[T_{\text{MAB}} \geq \frac{1}{2}pT_{\text{CMAB}}\right] \geq \frac{1}{2}$.*

Proof. $T_{\text{MAB}} = \sum_{t=1}^{T_{\text{CMAB}}} \gamma_t$. By multiplicative Chernoff bound (Fact 2),

$$\Pr[T_{\text{MAB}} \geq \frac{1}{2}pT_{\text{CMAB}}] \geq 1 - \left(\frac{e^{-\frac{1}{2}}}{\left(\frac{1}{2}\right)^{\frac{1}{2}}}\right)^{pT_{\text{CMAB}}} \geq \frac{1}{2},$$

when $pT_{\text{CMAB}} \geq 6$.

$$\Pr[T_{\text{MAB}} \geq \frac{1}{2}pT_{\text{CMAB}}] \geq 1 - \left(e^{-\frac{1}{8}pT_{\text{CMAB}}}\right) \geq \frac{1}{2},$$

when $pT_{\text{CMAB}} \geq 6$.

In the following, we overload the notation $\mathcal{D}$ to also represent a probabilistic distribution of the environment instance (a.k.a. outcome distribution) D, and use $D \sim \mathcal{D}$ to represent a random environment instance D drawn from the distribution $\mathcal{D}$.

Lemma 8. *Consider a random MAB environment instance D drawn from a distribution $\mathcal{D}$. Assume we have a lower bound $L(T_{\text{MAB}})$ of expected regret, i.e. for every natural number T_{MAB}, any MAB algorithm A has expected regret*

$$\mathop{\mathbb{E}}_{D\sim\mathcal{D}}[Reg^{A}_{\text{MAB},D}(T_{\text{MAB}})] \geq L(T_{\text{MAB}}).$$

Then consider the corresponding CMAB-T environment instance D. For every natural number $T_{\text{CMAB}} \geq 5p^{-1}$, any CMAB-T algorithm A has expected regret

$$\mathop{\mathbb{E}}_{D\sim\mathcal{D}}[Reg^{A}_{\text{CMAB},D}(T_{\text{CMAB}})] \geq \frac{1}{2}p^{-1}L(\frac{1}{2}pT_{\text{CMAB}}). \tag{39}$$

Proof. Without loss of generality, we may assume $L(T)$ is non-decreasing, as regret of any strategy increases as T increases.

We prove the lemma using the reduction described above. We run Algorithm 2 with A be the CMAB-T oracle and D be the environment instance. Let γ be the vector $(\gamma_1, \gamma_2, \ldots, \gamma_{T_{\text{CMAB}}})$. Every possible value of γ parameterizes Algorithm 2 into an algorithm plays MAB problem for $T_{\text{MAB}} = \sum_{t=1}^{T_{\text{CMAB}}} \gamma_t$ rounds. We denote this MAB algorithm with A_γ. By our assumption, $\mathbb{E}_{D\sim\mathcal{D}}[Reg^{A_\gamma}_{\text{MAB},D}(T_{\text{MAB}})] \geq L(T_{\text{MAB}})$.

Then we compare the regret in both cases. For a given distribution D, let $\mu_{i,D} = \mathbb{E}_{X\sim D}[X_i]$ and $\mu^*_D = \max_i \mu_{i,D}$. For MAB problem and every γ,

$$\begin{aligned}
\mathop{\mathbb{E}}_{D\sim\mathcal{D}}[Reg^{A_\gamma}_{\text{MAB},D}(T_{\text{MAB}})] &= \mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[T_{\text{MAB}} \cdot \mu^*_D - \mathbb{E}\left[\sum_{t=1}^{T_{\text{CMAB}}} \gamma_t X_{i_t}\right]\right] \\
&= \mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[\mathbb{E}\left[\sum_{t=1}^{T_{\text{CMAB}}} \gamma_t(\mu^*_D - X_{i_t})\right]\right] \\
&= \mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[\mathbb{E}\left[\sum_{t=1}^{T_{\text{CMAB}}} \gamma_t(\mu^*_D - \mu_{i_t,D})\right]\right],
\end{aligned}$$

where the inner expectation is taken over the rest randomness, including the randomness of i_t, which is based on the random feedback history and the possible randomness of algorithm A_γ. For CMAB-T, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{D\sim\mathcal{D}}[Reg^A_{\mathrm{CMAB},D}(T_{\mathrm{CMAB}})] \\
&= \mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[T_{\mathrm{CMAB}}\cdot\mu^*_D - \mathop{\mathbb{E}}_{\gamma\sim B_p^{T_{\mathrm{CMAB}}}}\left[\mathbb{E}\left[\sum_{t=1}^{T_{\mathrm{CMAB}}}\gamma_t p^{-1}X_{i_t}\right]\right]\right] \\
&= \mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[T_{\mathrm{CMAB}}\cdot\mu^*_D - \mathop{\mathbb{E}}_{\gamma\sim B_p^{T_{\mathrm{CMAB}}}}\left[\mathbb{E}\left[\sum_{t=1}^{T_{\mathrm{CMAB}}}\gamma_t p^{-1}\mu_{i_t,D}\right]\right]\right] \\
&= \mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[pT_{\mathrm{CMAB}}\cdot p^{-1}\mu^*_D - \mathop{\mathbb{E}}_{\gamma\sim B_p^{T_{\mathrm{CMAB}}}}\left[\mathbb{E}\left[\sum_{t=1}^{T_{\mathrm{CMAB}}}\gamma_t p^{-1}\mu_{i_t,D}\right]\right]\right] \\
&= \mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[\mathop{\mathbb{E}}_{\gamma\sim B_p^{T_{\mathrm{CMAB}}}}\left[\sum_{t=1}^{T_{\mathrm{CMAB}}}\gamma_t p^{-1}\mu^*_D\right] - \mathop{\mathbb{E}}_{\gamma\sim B_p^{T_{\mathrm{CMAB}}}}\left[\mathbb{E}\left[\sum_{t=1}^{T_{\mathrm{CMAB}}}\gamma_t p^{-1}\mu_{i_t,D}\right]\right]\right] \\
&= p^{-1}\mathop{\mathbb{E}}_{D\sim\mathcal{D},\gamma\sim B_p^{T_{\mathrm{CMAB}}}}\left[\mathbb{E}\left[\sum_{t=1}^{T_{\mathrm{CMAB}}}\gamma_t(\mu^* - \mu_{i_t,D})\right]\right],
\end{aligned}$$

where the innermost expectation is taken over the rest randomness such as the randomness of i_t. Therefore

$$\mathop{\mathbb{E}}_{D\sim\mathcal{D}}[Reg^A_{\mathrm{CMAB},D}(T_{\mathrm{CMAB}})] = p^{-1}\mathop{\mathbb{E}}_{D\sim\mathcal{D},\gamma\sim B_p^{T_{\mathrm{CMAB}}}}[Reg^{A_\gamma}_{\mathrm{MAB},D}(T_{\mathrm{MAB}})].$$

Calculation above also shows $\mathbb{E}_{D\sim\mathcal{D}}[Reg^{A_\gamma}_{\mathrm{MAB},D}(T_{\mathrm{MAB}})] \geq 0$. And by monotonicity of $L(T)$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{D}[Reg^A_{\mathrm{CMAB},D}(T_{\mathrm{CMAB}})] &= p^{-1}\mathop{\mathbb{E}}_{D,\gamma}[Reg^{A_\gamma}_{\mathrm{MAB},D}(T_{\mathrm{MAB}})] \\
&\geq p^{-1}\mathop{\mathbb{E}}_{D,\gamma}[\mathbb{I}\{T_{\mathrm{MAB}} \geq \frac{1}{2}pT_{\mathrm{CMAB}}\}Reg^{A_\gamma}_{\mathrm{MAB},D}(T_{\mathrm{MAB}})] \\
&\geq p^{-1}\mathop{\mathbb{E}}_{D,\gamma}[\mathbb{I}\{T_{\mathrm{MAB}} \geq \frac{1}{2}pT_{\mathrm{CMAB}}\}L(\frac{1}{2}pT_{\mathrm{CMAB}})] \\
&= p^{-1}\Pr_{D,\gamma}\{T_{\mathrm{MAB}} \geq \frac{1}{2}pT_{\mathrm{CMAB}}\}L(\frac{1}{2}pT_{\mathrm{CMAB}}) \\
&\geq \frac{1}{2}p^{-1}L(\frac{1}{2}pT_{\mathrm{CMAB}}).
\end{aligned}$$

□

Lemma 9. *Let m be the number of arms and T be the number of rounds. Let $\varepsilon = \frac{1}{10}\sqrt{m/T}$. Then define the family of MAB outcome distributions $\mathcal{D} = \{D_1, \ldots, D_m\}$ with*

$$\Pr_{D_j}\{X_i = 1\} = \begin{cases} \frac{1}{2} & \text{if } i \neq j \\ \frac{1}{2} + \varepsilon & \text{if } i = j \end{cases}.$$

Let D be a random environment instance uniformly drawn from $\mathcal{D}$, then for any MAB algorithm A,

$$\mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[Reg^A_{\mathrm{MAB},D}(T)\right] \geq \frac{\varepsilon T}{6} = \frac{1}{60}\sqrt{mT}.$$

Proof of Theorem 2. Let $\mathcal{D}$ be the family of outcome distributions defined in Lemma 9, and D is uniformly drawn from $\mathcal{D}$. Applying the result of Lemma 9 to Lemma 8, with $L(T) = \frac{1}{60}\sqrt{mT}$ in Lemma 8, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{D\sim\mathcal{D}}\left[Reg^A_{\mathrm{CMAB},D}(T)\right] &\geq \frac{1}{2}p^{-1}L(\frac{1}{2}pT) \\
&= \frac{1}{2}p^{-1}\cdot\frac{1}{60}\sqrt{\frac{1}{2}mpT} \\
&> \frac{1}{170}\sqrt{\frac{mT}{p}}.
\end{aligned}$$

Since D is uniformly drawn from $\mathcal{D}$, then there must exists a $D \in \mathcal{D}$ such that

$$Reg^{A}_{\text{CMAB},D}(T) \geq \frac{1}{170}\sqrt{\frac{mT}{p}}. \qquad \square$$

It is easy to show corresponding CMAB-T problem satisfies original bounded smoothness (Condition 5) with $f(x) = x$. So the theorem above gives an example that the upper bound in [7] is tight up to a $O(\sqrt{\log T})$ factor.

E.2 Proof of Theorem 3

Proof of Theorem 3. We regard this kind of CMAB-T problem instances as a variant of classical MAB, that each arm gives three possible outcomes, 0, 1, and $\perp$. Denote these arms with random variables $X'_1, \ldots, X'_n$. The reward is p^{-1} times of the outcome if the outcome is 0 or 1, while the reward is 0 if the outcome is $\perp$. This variant is equivalent to the CMAB-T instances: Outcome $X'_i = \perp$ corresponds to Bernoulli base arm X_i in CMAB-T not being triggered, outcome $X'_i = 1$ or 0 corresponds to Bernoulli base arm X_i being triggered and $X_i = 1$ or 0, respectively. Thus $\Pr[X'_i = \perp] = 1 - p$, $\Pr[X'_i = 0] = p(1 - \mu_i)$, and $\Pr[X'_i = 1] = p\mu_i$, where p is the triggering probability and μ_i is the expectation of X_i.

Let X and Y be random variables whose values are in the same finite set V. Define the KL-divergence

$$\mathrm{kl}(X, Y) = \sum_{x \in V} \Pr\{X = x\} \ln \frac{\Pr\{X = x\}}{\Pr\{Y = x\}}.$$

For example the KL-divergence between X'_1 and X'_2 is

$$\begin{aligned}
\mathrm{kl}(X'_1, X'_2) &= \Pr\{X'_1 = \perp\} \ln \frac{\Pr\{X'_1 = \perp\}}{\Pr\{X'_2 = \perp\}} + \Pr\{X'_1 = 0\} \ln \frac{\Pr\{X'_1 = 0\}}{\Pr\{X'_2 = 0\}} \\
&\quad + \Pr\{X'_1 = 1\} \ln \frac{\Pr\{X'_1 = 1\}}{\Pr\{X'_2 = 1\}} \\
&= (1 - p) \ln \frac{1 - p}{1 - p} + p(1 - \mu_1) \ln \frac{p(1 - \mu_1)}{p(1 - \mu_2)} + p\mu_1 \ln \frac{p\mu_1}{p\mu_2} \\
&= 0 + p(1 - \mu_1) \ln \frac{1 - \mu_1}{1 - \mu_2} + p\mu_1 \ln \frac{\mu_1}{\mu_2} \\
&= p \cdot \left[(1 - \mu_1) \ln \frac{1 - \mu_1}{1 - \mu_2} + \mu_1 \ln \frac{\mu_1}{\mu_2} \right] \\
&= p \cdot \mathrm{kl}(X_1, X_2). \qquad \square
\end{aligned}$$

Thus, intuitively it takes p^{-1} times more rounds to differentiate X'_1 and X'_2 than X_1 and X_2, which is stated formally in theorem below.

Proof. The analysis is generalized from the case that the arms are Bernoulli random variables. For an arm i, we use $N_i(T)$ to denote the number of times the arm i is played in T rounds. For each non-optimal arm i, i.e. $\mu_i < \mu^* < 1$, we show

$$\liminf_{T \to +\infty} \frac{\mathbb{E}[N_i(T)]}{\ln T} \geq \frac{p^{-1}}{\mathrm{kl}(X_i, X_{i^*})} = \frac{1}{\mathrm{kl}(X'_i, X'_{i^*})}. \tag{40}$$

Then by formula

$$Reg^{A}_{\boldsymbol{\mu}}(T) = \sum_{i:\mu_i < \mu^*} \mathbb{E}[N_i(T)]\Delta_i,$$

the theorem holds.

Without loss of generality, we may assume arm 1 is an optimal arm and arm 2 is non-optimal. We prove Eq. (40) for arm 2 and then the inequality holds for every arm. Consider that if we replace arm 2 with a fictional arm $2'$, which has an expectation $\mu_{2'}$ slightly greater than μ_1, then arm 1 will

become non-optimal and strategy A will play arm 1 for $o(n^a)$ times for any $a > 0$. So strategy A must play arm 2 for enough times, to differentiate from arm $2'$.

Formally, let $\varepsilon > 0$ be any positive real number. Let $\mu_{2'}$ be a real number such that $\mu_{2'} > \mu_1$ and

$$\mathrm{kl}(X_2, X_{2'}) = (1-\mu_2)\ln\frac{1-\mu_2}{1-\mu_{2'}} + \mu_2 \ln\frac{\mu_2}{\mu_{2'}} < (1+\varepsilon)\mathrm{kl}(X_2, X_1). \tag{41}$$

There exists such $\mu_{2'}$, because the left hand side of (41) is continuous as a function of $\mu_{2'}$. We use $\mathbb{E}'$ and $\Pr'$ to denote expectation and probability in the circumstance that arm X_2 is replaced by arm $X_{2'}$.

We define the empirical KL-divergence after the first s samples of the arm 2/2$'$,

$$\widehat{\mathrm{kl}}_s = \sum_{t=1}^{s} Y_t,$$

where

$$Y_t = \begin{cases} \ln\frac{1-\mu_2}{1-\mu_{2'}}, & \text{if } X'_{2,t} = 0, \\ \ln\frac{\mu_2}{\mu_{2'}}, & \text{if } X'_{2,t} = 1, \\ 0, & \text{if } X'_{2,t} = \perp . \end{cases}$$

and $X'_{2,t}$ is result of the t-th sample of arm 2/2$'$. Note that (Y_t) are independent and $\mathbb{E}[Y_t] = \mathrm{kl}(X'_2, X'_{2'})$.

First we prove

$$\Pr\left\{N_2(T) < \frac{1-\varepsilon}{\mathrm{kl}(X'_2, X'_{2'})}\ln T \wedge \widehat{\mathrm{kl}}_{N_2(T)} \le \left(1-\frac{\varepsilon}{2}\right)\ln T\right\} = o(1). \tag{42}$$

We use the shorthands

$$C_T = \left\{N_2(T) < \frac{1-\varepsilon}{\mathrm{kl}(X'_2, X'_{2'})}\ln T \wedge \widehat{\mathrm{kl}}_{N_2(T)} \le \left(1-\frac{\varepsilon}{2}\right)\ln T\right\}, \tag{43}$$

and

$$f_T = \frac{1-\varepsilon}{\mathrm{kl}(X'_2, X'_{2'})}\ln T.$$

If arm 2 is replaced by arm $2'$, we have

$$\Pr'\{C_T\} \le \Pr'\{N_2(T) < f_T\} \le \frac{\mathbb{E}'[T - N_2(T)]}{T - f_T},$$

where the second inequality is due to Markov's inequality. Recall the definition of consistent strategy, as $2'$ is the only optimal arm, we have $\mathbb{E}'[T - N_2(T)] = o(T^{\frac{\varepsilon}{2}})$. And by $T - f_T = \Omega(T)$, $\Pr'\{C_T\} = o(T^{\frac{\varepsilon}{2}-1})$. Then we use the property of KL-divergence

$$\Pr\{C_T\} = \mathbb{E}'\left[\mathbb{I}\{C_T\}\cdot\exp\left(\widehat{\mathrm{kl}}_{N_2(T)}\right)\right],$$

then

$$\Pr\{C_T\} = \mathbb{E}'\left[\mathbb{I}\{C_n\}\cdot\exp\left(\widehat{\mathrm{kl}}_{N_2(T)}\right)\right] \le \Pr'\{C_T\}\cdot\exp\left[\left(1-\frac{\varepsilon}{2}\right)\ln T\right] = \Pr'\{C_T\}\cdot T^{1-\frac{\varepsilon}{2}} = o(1).$$

Second, we prove

$$\Pr\left\{N_2(T) < f_T \wedge \widehat{\mathrm{kl}}_{T_2(T)} > \left(1-\frac{\varepsilon}{2}\right)\ln T\right\} = o(1). \tag{44}$$

We have

$$\begin{aligned}\Pr\left\{N_2(T) < f_T \wedge \widehat{\mathrm{kl}}_{N_2(T)} > \left(1-\frac{\varepsilon}{2}\right)\ln T\right\} &\le \Pr\left\{N_2(T) < f_T \wedge \max_{s\le f_T}\widehat{\mathrm{kl}}_s > \left(1-\frac{\varepsilon}{2}\right)\ln T\right\} \\ &\le \Pr\left\{\max_{s\le f_T}\widehat{\mathrm{kl}}_s > \left(1-\frac{\varepsilon}{2}\right)\ln T\right\}.\end{aligned}$$

Recall the definition of $\widehat{\mathrm{kl}}_s$, which is a summation of independent random variables with the same distribution over a finite support, whose expectation is $\mathrm{kl}(X_2', X_{2'}')$. So we apply the maximal version of the strong law of large numbers, and then (44) holds, as $f_T \cdot \mathrm{kl}(X_2', X_{2'}') = (1-\varepsilon)\ln T$.

In conclusion, combining Eq. (42) and (44), we have $\Pr\{N_2(T) < f_T\} = o(1)$, implying

$$\begin{aligned}\mathbb{E}\left[N_2(T)\right] &\geq (1-o(1)) \cdot f_T \\ &= (1-o(1)) \cdot \frac{1-\varepsilon}{\mathrm{kl}(X_2', X_{2'}')} \ln T \\ &\geq (1-o(1)) \cdot \frac{1-\varepsilon}{1+\varepsilon} \frac{\ln T}{\mathrm{kl}(X_2', X_1')}.\end{aligned}$$

Then (40) holds, as ε can be any positive real number, and thus the theorem holds. □

F Results with ∞-norm TPM Conditions

F.1 TPM Conditions with the ∞-norm

We first restate the original bounded smoothness condition in [7] below, which is an ∞-norm based condition.

Condition 5 (Bounded Smoothness). *We say that a CMAB-T problem instance satisfies* bounded smoothness*, if there exists a continuous, strictly increasing (and thus invertible) function $f(\cdot)$ with $f(0) = 0$, such that for any two distributions $D, D' \in \mathcal{D}$ with expectation vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)$ and $\boldsymbol{\mu}' = (\mu_1', \ldots, \mu_m')$, and for any $\Lambda > 0$, we have $|r_{\boldsymbol{\mu}}(S) - r_{\boldsymbol{\mu}'}(S)| \leq f(\Lambda)$ if $\max_{i \in \tilde{S}} |\mu_i - \mu_i'| \leq \Lambda$, for all $S \in \mathcal{S}$, where $\tilde{S} = \{i \in [m] \mid \Pr_{X \sim D, \tau}\{i \in \tau(S, X)\} > 0\}$ is the set of arms that could be triggered by action S.*

Note that $f(\cdot)$ may depend on problem instance parameters such as m, but not on action S or mean vectors $\boldsymbol{\mu}$, $\boldsymbol{\mu}'$.

Similar to the 1-norm case, we use triggering probabilities to modulate the bounded smoothness condition to obtain the following TPM version:

Condition 6. (∞-Norm TPM Bounded Smoothness) *We say a CMAB-T problem instance satisfies the* triggering-probability-modulated (TPM) bounded smoothness *with bounded smoothness function $f(x)$, if for any two distributions $D, D' \in \mathcal{D}$ with expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, any action S and any $\Lambda > 0$, we have $|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \leq f(\Lambda)$ if $\max_{i \in [m]} p_i^{D,S} |\mu_i - \mu_i'| \leq \Lambda$.*

Note that Condition 6 is stronger than Condition 5 under the same bounded smoothness function f. This is because if we have $\max_{i \in [m]} |\mu_i - \mu_i'| \leq \Lambda$, then we have $\max_{i \in [m]} p_i^{D,S} |\mu_i - \mu_i'| \leq \Lambda$. Then if Condition 6 holds, we have $|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \leq f(\Lambda)$. This means that if Condition 6 holds, we have $|r_S(\boldsymbol{\mu}) - r_S(\boldsymbol{\mu}')| \leq f(\Lambda)$ if $\max_{i \in [m]} |\mu_i - \mu_i'| \leq \Lambda$, which is exactly Condition 5.

F.2 Theorem and Proofs with ∞-norm TPM Conditions

Theorem 5. *Suppose a CMAB-T problem instance $([m], \mathcal{S}, \mathcal{D}, D^{\mathrm{trig}}, R)$ satisfies monotonicity (Condition 1). For a fixed environment instance $D \in \mathcal{D}$ with expectation vector $\boldsymbol{\mu}$, the T-round (α, β)-approximation regret bound using an (α, β)-approximation oracle in various cases are given below.*

(1) For the CUCB algorithm on a problem instance that satisfies TPM bounded smoothness (Condition 6) with bounded smoothness function $f(x)$, together with $\Delta_{\min} > 0$, the regret is at most

$$\sum_{i \in [m]} 78 \ln T \left(\frac{\Delta_{\min}^i}{f^{-1}(\Delta_{\min}^i)^2} + \int_{\Delta_{\min}^i}^{\Delta_{\max}^i} \frac{1}{f^{-1}(x)^2} \, \mathrm{d}x \right) + m \cdot \left[\left(\frac{\pi^2}{6} + 1 \right) \lceil -\log_2 f^{-1}(\Delta_{\min}) \rceil_0 + \frac{\pi^2}{3} + 1 \right] \cdot \Delta_{\max};$$

(2) For the CUCB algorithm on a problem instance that satisfies TPM bounded smoothness (Condition 6) with bounded smoothness function $f(x) = ax$, the regret is at most

$$25a\sqrt{mT\ln T} + m \cdot \left[\left(\frac{\pi^2}{6}+1\right)\left\lceil -\log_2(\sqrt{156m\ln T/T})\right\rceil_0 + \frac{\pi^2}{3}+1\right]\cdot\Delta_{\max};$$

We have several remarks on Theorem 5. First, the condition $\Delta_{\min} > 0$ automatically holds if the action space $\mathcal{S}$ is finite. Thus it is not an extra condition comparing to the result in [7] when actions are set of base arms. If $\Delta_{\min}$ is zero due to infinite $\mathcal{S}$, then we do not have regret bound as in (1), but we still have regret bound as in (2). Second, the regret bound in (1) is distribution-dependent bound, since it depends on $\Delta^i_{\min}$, which is determined by the distribution D; regret bounds in (2) is distribution-independent bound, since $\Delta_{\max}$ can be easily replaced by a quantity only depending on the problem instance, such as the maximum possible reward value. Third, when $\Delta^i_{\min} = +\infty$, $\frac{\Delta^i_{\min}}{f^{-1}(\Delta^i_{\min})^2} = 0$.

F.2.1 Proof of Theorem 5

In this subsection, we focus on giving a roadmap to prove Theorem 5 and showing the new techniques we invented to improve the regret bound. The remaining part of the proof is roughly the new calculation based on the old techniques (c.f. [7]).

In this subsection, we omit (α, β)-approximation for clarity, in other words, we assume $\alpha = \beta = 1$. Generalization to accommodate (α, β) approximation can be found in the discussion section.

To exploit the advantage of TPM bounded smoothness condition (Conditions 6), for each arm i, we divide actions into groups according to $p_i^{D,S}$.

For convenience, we also allow to index the counters with $q_i^{D,S_t} > 0$, such that $N_{i,q_i^{D,S_t}}$ indicates the same counter as $N_{i,j}$ with $q_i^{D,S_t} = 2^{-j}$.

We use a shorthand as follows. For every arm i and action S, define

$$q_i^{D,S} = \begin{cases} 2^{-j}, & \text{if } S \in \mathcal{S}^D_{i,j}, \\ 0, & \text{if } p_i^{D,S} = 0. \end{cases}$$

Definition 8.

$$\ell_t(\Delta, q) = \begin{cases} 0, & \textit{if } q \le \frac{1}{2}f^{-1}(\Delta), \\ \lfloor \frac{6\ln t}{f^{-1}(\Delta)^2} \rfloor + 1, & \textit{if } q = 1, \\ \lfloor \frac{72q\ln t}{f^{-1}(\Delta)^2} \rfloor + 1, & \textit{otherwise}. \end{cases}$$

To unify the proofs for distribution-dependent and distribution-independent bounds, we introduce a positive real number M. To prove the distribution-dependent bound, we will let $M = \Delta_{\min}$ or $M = \Delta^i_{\min}$ in some circumstances. To prove the distribution-independent bound, we will let $M = \tilde{\Theta}(T^{-1/2})$ to balance bounds for $Reg(\{\Delta_{S_t} \ge M\})$ and $Reg(\{\Delta_{S_t} < M\})$. And we implement $\mathcal{N}^{\mathrm{t}}_t$ (Definition 7) with $j^i_{\max} = j_{\max}(M) = \lceil -\log_2 f^{-1}(M)\rceil_0$ The following are three technical claims used in the main proof, and we define the proofs of these claims to Section F.2.2.

Claim 1 (Bound of insufficiently sampled regret). *For any CMAB-T problem instance, any bounded smoothness function $f(x)$, any algorithm, any arm i, any natural number j and any positive real number M,*

$$Reg(\{\Delta_{S_t} \ge M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) \le \ell_T(M, 2^{-j})M + \int_M^{\max\{\Delta^i_{\max}, M\}} \ell_T(x, 2^{-j})\,\mathrm{d}x.$$

Claim 2 (Bound of sufficiently sampled regret for CUCB). *For the CUCB algorithm on a problem instance that satisfies TPM bounded smoothness (Condition 6) with bounded smoothness function $f(x)$,*

$$Reg(\{\Delta_{S_t} \ge M, \forall i, N_{i,q_i^{S_t},t-1} \ge \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \le m \cdot (\lceil -\log_2 f^{-1}(M)\rceil_0 + 2) \cdot \frac{\pi^2}{6} \cdot \Delta_{\max}.$$

We continue the proof of Theorem 5. Fix a value $M > 0$, we have

$$\begin{aligned}
Reg(\{\}) &= Reg(\{\Delta_{S_t} < M\}) + Reg(\{\Delta_{S_t} \geq M\}) \\
&= Reg(\{\Delta_{S_t} < M\}) + Reg(\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&\quad + Reg(\{\Delta_{S_t} \geq M, \exists i, N_{i,q_i^{S_t},t-1} < \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&\leq Reg(\{\Delta_{S_t} < M\}) + Reg(\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&\quad + \sum_{i\in[m]} Reg(\{\Delta_{S_t} \geq M, N_{i,q_i^{S_t},t-1} < \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&\leq Reg(\{\Delta_{S_t} < M\}) + Reg(\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&\quad + \sum_{i\in[m]} \sum_{j\geq 0} Reg(\{\Delta_{S_t} \geq M, S_t \in \mathcal{S}_{i,j}, N_{i,q_i^{S_t},t-1} < \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&= Reg(\{\Delta_{S_t} < M\}) + Reg(\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&\quad + \sum_{i\in[m]} \sum_{j\geq 0} Reg(\{\Delta_{S_t} \geq M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}). \qquad (45)
\end{aligned}$$

For the last part, if $j \geq \lceil -\log_2 f^{-1}(M)\rceil_0 + 1$, then $2^{-j} \leq \frac{1}{2} f^{-1}(M)$ and

$$\frac{1}{2} f^{-1}(\Delta_{S_t}) \geq \frac{1}{2} f^{-1}(M) \geq 2^{-j}.$$

By Definition 8, $\ell_T(\Delta_{S_t}, 2^{-j}) = 0$. Then $N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})$ is impossible, so

$$\sum_{j\geq \lceil -\log_2 f^{-1}(M)\rceil_0 + 1} Reg(\{\Delta_{S_t} \geq M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) = 0.$$

Lemma 10. *For every arm i, the event-filtered regret*

$$\begin{aligned}
&\sum_{j\geq 0} Reg(\{\Delta_{S_t} \geq M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) \qquad (46) \\
&\leq 78 \ln T \left(\frac{M}{f^{-1}(M)^2} + \int_M^{\max\{\Delta^i_{\max}, M\}} \frac{1}{f^{-1}(x)^2}\, \mathrm{d}x \right) + (j_{\max}(M) + 1) \cdot \Delta^i_{\max}.
\end{aligned}$$

Proof. If $M > \Delta^i_{\max}$, it is impossible to have $\Delta_{S_t} \geq M$ and $S_t \in \mathcal{S}_{i,j}$ at the same time and then (46) $= 0$. Then the lemma holds trivially. So we may assume that $M \leq \Delta^i_{\max}$. By Claim 1,

$$\begin{aligned}
(46) &= \sum_{j=0}^{j_{\max}(M)} Reg(\{\Delta_{S_t} \geq M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) \\
&\leq \sum_{j=0}^{j_{\max}(M)} \left(\ell_T(M, 2^{-j}) M + \int_M^{\max\{\Delta^i_{\max}, M\}} \ell_T(x, 2^{-j})\, \mathrm{d}x \right) \\
&= \sum_{j=0}^{j_{\max}(M)} \left(\ell_T(M, 2^{-j}) M + \int_M^{\Delta^i_{\max}} \ell_T(x, 2^{-j})\, \mathrm{d}x \right) \\
&= \sum_{j=0}^{j_{\max}(M)} \ell_T(M, 2^{-j}) M + \int_M^{\Delta^i_{\max}} \sum_{j=0}^{j_{\max}(M)} \ell_T(x, 2^{-j})\, \mathrm{d}x. \qquad (47)
\end{aligned}$$

We then expand the notation $\ell_T(\Delta, q)$ (c.f. Definition 8) with

$$\ell_T(\Delta, q) \leq \begin{cases} \frac{6 \ln T}{f^{-1}(\Delta)^2} + 1, & \text{if } q = 1, \\ \frac{72 q \ln T}{f^{-1}(\Delta)^2} + 1, & \text{otherwise.} \end{cases}$$

So for any $x \in [M, \Delta^i_{\max}]$,

$$\begin{aligned}
\sum_{j=0}^{j_{\max}(M)} \ell_T(x, 2^{-j}) &= \ell_T(x,1) + \sum_{j=1}^{j_{\max}(M)} \ell_T(x, 2^{-j}) \\
&\le \left(\frac{6\ln T}{f^{-1}(x)^2} + 1\right) + \sum_{j=1}^{j_{\max}(M)} \left(\frac{72\cdot 2^{-j}\ln T}{f^{-1}(x)^2} + 1\right) \\
&= \frac{6\ln T}{f^{-1}(x)^2} + \sum_{j=1}^{j_{\max}(M)} \frac{72\cdot 2^{-j}\ln T}{f^{-1}(x)^2} + j_{\max}(M) + 1 \\
&\le \frac{6\ln T}{f^{-1}(x)^2} + \frac{72\ln T}{f^{-1}(x)^2} + j_{\max}(M) + 1 \\
&= \frac{78\ln T}{f^{-1}(x)^2} + j_{\max}(M) + 1.
\end{aligned}$$

Then we continue (47) with

$$\begin{aligned}
(47) &\le \left(\frac{78\ln T}{f^{-1}(M)^2} + j_{\max}(M) + 1\right)\cdot M + \int_M^{\Delta^i_{\max}} \left(\frac{78\ln T}{f^{-1}(x)^2} + j_{\max}(M) + 1\right) \mathrm{d}x \\
&= \frac{78\ln T}{f^{-1}(M)^2}\cdot M + \int_M^{\Delta^i_{\max}} \frac{78\ln T}{f^{-1}(x)^2}\,\mathrm{d}x + (j_{\max}(M)+1)\cdot\Delta^i_{\max} \\
&= 78\ln T\left(\frac{M}{f^{-1}(M)^2} + \int_M^{\Delta^i_{\max}} \frac{1}{f^{-1}(x)^2}\,\mathrm{d}x\right) + (j_{\max}(M)+1)\cdot\Delta^i_{\max}.
\end{aligned}$$

Hence the lemma holds. □

Lemma 11. *For event-filtered regret*

$$Reg(\{\Delta_{S_t} < M\}) + \sum_{i\in[m]}\sum_{j\ge 0} Reg(\{\Delta_{S_t} \ge M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}), \tag{48}$$

(1) take $M = \Delta_{\min}$ when $\Delta_{\min} > 0$,

$$(48) \le \sum_{i\in[m]} 78\ln T\left(\frac{\Delta^i_{\min}}{f^{-1}(\Delta^i_{\min})^2} + \int_{\Delta^i_{\min}}^{\Delta^i_{\max}} \frac{1}{f^{-1}(x)^2}\,\mathrm{d}x\right) + m\cdot(j_{\max}(\Delta_{\min})+1)\cdot\Delta_{\max};$$

(2) if $f(x) = ax$, then take $M = a\sqrt{156m\ln T/T}$,

$$(48) < 25a\sqrt{mT\ln T} + m\cdot(j_{\max}(a\sqrt{156m\ln T/T}) + 1)\cdot\Delta_{\max}.$$

Proof. (1) If $\Delta_{S_t} < M = \Delta_{\min}$, then $\Delta_{S_t} = 0$. So $Reg(\{\Delta_{S_t} < M\}) \le 0$. For every $i \in [m]$ and every integer j, we may replace M with $\Delta^i_{\min}$ as below.

$$\begin{aligned}
&Reg(\{\Delta_{S_t} \ge M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) \\
&= Reg(\{\Delta_{S_t} \ge \Delta_{\min}, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) \\
&= Reg(\{\Delta_{S_t} \ge \Delta^i_{\min}, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}).
\end{aligned} \tag{49}$$

Then apply Lemma 10 with $M = \Delta^i_{\min}$, we have

$$\begin{aligned}
(48) &= \sum_{i\in[m]}\sum_{j\ge 0} Reg(\{\Delta_{S_t} \ge M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) \\
&\le \sum_{i\in[m]}\left[78\ln T\left(\frac{\Delta^i_{\min}}{f^{-1}(\Delta^i_{\min})^2} + \int_{\Delta^i_{\min}}^{\Delta^i_{\max}} \frac{1}{f^{-1}(x)^2}\,\mathrm{d}x\right) + (j_{\max}(\Delta^i_{\min})+1)\cdot\Delta^i_{\max}\right] \\
&\le \sum_{i\in[m]} 78\ln T\left(\frac{\Delta^i_{\min}}{f^{-1}(\Delta^i_{\min})^2} + \int_{\Delta^i_{\min}}^{\Delta^i_{\max}} \frac{1}{f^{-1}(x)^2}\,\mathrm{d}x\right) + m\cdot(j_{\max}(\Delta_{\min})+1)\cdot\Delta_{\max}.
\end{aligned}$$

(2) By Lemma 10, for every arm i,

$$\begin{aligned}
&\sum_{j\geq 0} Reg(\{\Delta_{S_t} \geq M, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(\Delta_{S_t}, 2^{-j})\}) \\
&\leq 78 \ln T \left(\frac{M}{f^{-1}(M)^2} + \int_M^{\Delta^i_{\max}} \frac{1}{f^{-1}(x)^2} \,\mathrm{d}x \right) + (j_{\max}(M)+1) \cdot \Delta^i_{\max} \\
&= 78 \ln T \left(\frac{M}{(a^{-1}M)^2} + \int_M^{\Delta^i_{\max}} \frac{1}{(a^{-1}x)^2} \,\mathrm{d}x \right) + (j_{\max}(M)+1) \cdot \Delta^i_{\max} \\
&= 78 \ln T \left(\frac{1}{a^{-2}M} + \int_M^{\Delta^i_{\max}} \frac{1}{a^{-2}x^2} \,\mathrm{d}x \right) + (j_{\max}(M)+1) \cdot \Delta^i_{\max} \\
&\leq 78 \ln T \left(\frac{1}{a^{-2}M} + \frac{1}{a^{-2}M} \right) + (j_{\max}(M)+1) \cdot \Delta^i_{\max} \\
&= \frac{156 \ln T}{a^{-2}M} + (j_{\max}(M)+1) \cdot \Delta_{\max}. \qquad (50)
\end{aligned}$$

$Reg(\{\Delta_{S_t} < M\}) < TM$ as the regret in each round is less than M. So by (50) and take $M = a\sqrt{156m \ln T/T}$,

$$\begin{aligned}
(48) &< TM + \frac{156m \ln T}{a^{-2}M} + m \cdot (j_{\max}(M)+1) \cdot \Delta_{\max} \\
&= a\sqrt{156mT \ln T} + a\sqrt{156mT \ln T} + m \cdot (j_{\max}(M)+1) \cdot \Delta_{\max} \\
&< 25a\sqrt{mT \ln T} + m \cdot (j_{\max}(a\sqrt{156m \ln T/T})+1) \cdot \Delta_{\max}. \qquad \square
\end{aligned}$$

Proof of Theorem 5. (1) Since $\Delta_{\min} > 0$, we can take $M = \Delta_{\min}$. By Lemma 11(1) and Claim 2, we continue Inequality (45) as below.

$$\begin{aligned}
(45) &\leq \sum_{i\in[m]} 78 \ln T \left(\frac{\Delta^i_{\min}}{f^{-1}(\Delta^i_{\min})^2} + \int_{\Delta^i_{\min}}^{\Delta^i_{\max}} \frac{1}{f^{-1}(x)^2} \,\mathrm{d}x \right) + m \cdot (j_{\max}(\Delta_{\min})+1) \cdot \Delta_{\max} \\
&\quad + m \cdot (j_{\max}(\Delta_{\min})+2) \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} \\
&= \sum_{i\in[m]} 78 \ln T \left(\frac{\Delta^i_{\min}}{f^{-1}(\Delta^i_{\min})^2} + \int_{\Delta^i_{\min}}^{\Delta^i_{\max}} \frac{1}{f^{-1}(x)^2} \,\mathrm{d}x \right) \\
&\quad + m \cdot \left[\left(\frac{\pi^2}{6} + 1 \right) \lceil -\log_2 f^{-1}(\Delta_{\min}) \rceil_0 + \frac{\pi^2}{3} + 1 \right] \cdot \Delta_{\max}.
\end{aligned}$$

(2) Take $M = a\sqrt{156m \ln T/T}$, by Lemma 11(2) and Claim 2, we continue Inequality (45) as below.

$$\begin{aligned}
(45) &\leq 25a\sqrt{mT \ln T} + m \cdot (j_{\max}(a\sqrt{156m \ln T/T})+1) \cdot \Delta_{\max} \\
&\quad + m \cdot (j_{\max}(a\sqrt{156m \ln T/T})+2) \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} \\
&= 25a\sqrt{mT \ln T} + m \cdot \left[\left(\frac{\pi^2}{6} + 1 \right) \left\lceil -\log_2(\sqrt{156m \ln T/T}) \right\rceil_0 + \frac{\pi^2}{3} + 1 \right] \cdot \Delta_{\max}. \square
\end{aligned}$$

F.2.2 Proof details

In this subsection, we finish the remaining part of the proof, i.e. the proofs of the claims. We first prove the bound of sufficiently sampled part, namely Claims 2. To do so, we define two kinds of niceness, that the difference between μ_i and $\hat{\mu}_i$ is small enough and that T_i is large enough comparing with $N_{i,j}$, and then show that both kinds of niceness are satisfied with high probability and if so, it is impossible to play a bad action. We then prove Claim 1. In this subsection we assume M is already

defined as a positive real number as in the proof of Theorem 5. Notations $\hat{\boldsymbol{\mu}}_t, \hat{\mu}_{i,t}, \bar{\boldsymbol{\mu}}_t, \bar{\mu}_{i,t}$ denote the values of $\hat{\boldsymbol{\mu}}, \hat{\mu}_i, \bar{\boldsymbol{\mu}}, \bar{\mu}_i$ at the end of round t, respectively.

We now prove the claims.

Proof of Claim 2. Explicitly,

$$\begin{aligned}
&Reg(\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}) \\
&= \sum_{t=1}^{T} \mathbb{E}[\Delta_{S_t} \cdot \mathbb{I}\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}] \\
&\leq \sum_{t=1}^{T} \Pr\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\} \cdot \Delta_{\max}. \qquad (51)
\end{aligned}$$

We only need to bound $\Pr\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}$, i.e. the probability that for every i, there is $N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})$, but an action S_t with $\Delta_{S_t} \geq M$ is still played. Let event $\mathcal{E}_t = \{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}$. We now prove the claim that event $\mathcal{E}_t$ is not empty only when $\neg(\mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}})$, or equivalently if both the sampling and triggering are nice at the beginning of round t, then event $\mathcal{E}_t$ is empty. If the sampling is nice at the beginning of round t, then

$$\bar{\mu}_{i,t-1} = \min\{\hat{\mu}_{i,t-1} + \rho_{i,t}, 1\} \geq \mu_i.$$

By monotonicity, $r_S(\bar{\boldsymbol{\mu}}_{t-1}) \geq r_S(\boldsymbol{\mu})$ for every action S, so $\mathrm{opt}_{\bar{\boldsymbol{\mu}}_{t-1}} \geq \mathrm{opt}_{\boldsymbol{\mu}}$. As action S_t is chosen by Oracle with input $\bar{\boldsymbol{\mu}}_{t-1}$, it must be that $r_{S_t}(\bar{\boldsymbol{\mu}}_{t-1}) = \mathrm{opt}_{\bar{\boldsymbol{\mu}}_{t-1}} \geq \mathrm{opt}_{\boldsymbol{\mu}}$, so $r_{S_t}(\bar{\boldsymbol{\mu}}_{t-1}) - r_{S_t}(\boldsymbol{\mu}) \geq \mathrm{opt}_{\boldsymbol{\mu}} - r_{S_t}(\boldsymbol{\mu}) = \Delta_{S_t}$. We are going to show the claim by assuming $\mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}}$ and showing $\forall i, p_i^{S_t} |\bar{\mu}_{i,t-1} - \mu_i| < f^{-1}(\Delta_{S_t})$, then by ∞-norm TPM bounded smoothness (Condition 6), $r_{S_t}(\bar{\boldsymbol{\mu}}_{t-1}) - r_{S_t}(\boldsymbol{\mu}) < \Delta_{S_t}$, which is a contradiction. Note that here we do need strict inequality "$<$" instead of "$\leq$" when applying Condition 6. This can be done because i has at most m choices and the bounded smoothness function f is continuous and strictly increasing, so we can use a small enough $\varepsilon > 0$ such that $\forall i, p_i^{S_t} |\bar{\mu}_{i,t-1} - \mu_i| \leq f^{-1}(\Delta_{S_t} - \varepsilon)$, and thus $r_{S_t}(\bar{\boldsymbol{\mu}}_{t-1}) - r_{S_t}(\boldsymbol{\mu}) \leq \Delta_{S_t} - \varepsilon < \Delta_{S_t}$.

Below we omit S_t from $\Delta_{S_t}, p_i^{S_t}$ and $q_i^{S_t}$. If $f^{-1}(\Delta) > p_i$, then $p_i |\bar{\mu}_{i,t-1} - \mu_i| \leq p_i |1 - 0| < f^{-1}(\Delta)$ without any dependency on sampling. If $f^{-1}(\Delta) \leq p_i$, then $q_i \leq 2^{\lceil -\log_2 f^{-1}(\Delta) \rceil} \leq 2^{j_{\max}(M)}$. When the sampling is nice (Definition 4), $\bar{\mu}_{i,t-1} \leq \hat{\mu}_{i,t-1} + \rho_{i,t} < \mu_i + 2\rho_{i,t}$. On the other hand, $|\bar{\mu}_{i,t-1} - \mu_i| \leq |1 - 0| = 1$. When the triggering is nice (Definition 7), if $\sqrt{\frac{6 \ln t}{\frac{1}{3} N_{i,q_i,t-1} \cdot q_i}} \leq 1$, then $2\rho_{i,t} \leq \sqrt{\frac{6 \ln t}{\frac{1}{3} N_{i,q_i,t-1} \cdot q_i}}$. So regardless whether $\sqrt{\frac{6 \ln t}{\frac{1}{3} N_{i,q_i,t-1} \cdot q_i}} \leq 1$, $|\bar{\mu}_{i,t-1} - \mu_i| \leq \sqrt{\frac{6 \ln t}{\frac{1}{3} N_{i,q_i,t-1} \cdot q_i}}$. Event $\mathcal{E}_t$ implies that $N_{i,q_i,t-1} \geq \ell_T(\Delta, q_i) \geq \ell_t(\Delta, q_i)$ (since $t \leq T$). So

$$\begin{aligned}
p_i |\bar{\mu}_{i,t-1} - \mu_i| &\leq p_i \sqrt{\frac{6 \ln t}{\frac{1}{3} N_{i,q_i,t-1} \cdot q_i}} \leq p_i \sqrt{\frac{6 \ln t}{\frac{1}{3} \ell_t(\Delta, q_i) \cdot q_i}} < p_i \sqrt{\frac{6 \ln t}{\frac{1}{3} \frac{72 q_i \ln t}{f^{-1}(\Delta)^2} \cdot q_i}} \\
&= p_i \sqrt{\frac{f^{-1}(\Delta)^2}{4 q_i^2}} \leq p_i \sqrt{\frac{f^{-1}(\Delta)^2}{p_i^2}} = f^{-1}(\Delta).
\end{aligned}$$

Hence, the claim holds.

The claim implies that $\Pr\{\mathcal{E}_t\} \leq \Pr\{\neg(\mathcal{N}_t^{\mathrm{s}} \wedge \mathcal{N}_t^{\mathrm{t}})\} \leq \Pr\{\neg \mathcal{N}_t^{\mathrm{s}}\} + \Pr\{\neg \mathcal{N}_t^{\mathrm{t}}\}$. By Lemmas 3 and 4, we have $\Pr\{\mathcal{E}\} \leq (2 + j_{\max}(M)) m t^{-2}$. Plugging it into Inequality (51), we have

$$\begin{aligned}
Reg(\{\Delta_{S_t} \geq M, \forall i, N_{i,q_i^{S_t},t-1} \geq \ell_T(\Delta_{S_t}, q_i^{S_t})\}) &\leq \sum_{t=1}^{T} (2 + j_{\max}(M)) m t^{-2} \cdot \Delta_{\max} \\
&\leq m \cdot (\lceil -\log_2 f^{-1}(M) \rceil_0 + 2) \cdot \frac{\pi^2}{6} \cdot \Delta_{\max} \square
\end{aligned}$$

Proof of Claim 1. Let x be any real number that $x \geq M > 0$. In any round when an action S with $S \in \mathcal{S}_{i,j}$ is played, $N_{i,j}$ is increased by 1. So

$$\sum_{t=1}^{T} \Pr\{S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(x, 2^{-j})\} \leq \ell_T(x, 2^{-j}).$$

If we add an additional restriction $\Delta_{S^t} \geq x$, the probability will not increase, so

$$\sum_{t=1}^{T} \Pr\{\Delta_{S_t} \geq x, S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(x, 2^{-j})\} \leq \ell_T(x, 2^{-j}).$$

We use the shorthand $\mathcal{E}_{i,j}^{S_t}$ to denote the event $\{S_t \in \mathcal{S}_{i,j}, N_{i,j,t-1} < \ell_T(x, 2^{-j})\}$. Suppose X is a non-negative random variable with $\Pr\{X \geq M\} = p$ and $\Pr\{X = 0\} = 1 - p$. Then by the basic principal on expectation, we have

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{+\infty} \Pr\{X \geq x\} \, \mathrm{d}x = \int_0^{M} \Pr\{X \geq x\} \, \mathrm{d}x + \int_M^{+\infty} \Pr\{X \geq x\} \, \mathrm{d}x \\
&= pM + \int_M^{+\infty} \Pr\{X \geq x\} \, \mathrm{d}x.
\end{aligned}$$

Applying the above, we have

$$\begin{aligned}
&Reg(\{\Delta_{S_t} \geq M\} \cap \mathcal{E}_{i,j}^{S_t}) \\
&= \sum_{t=1}^{T} \mathbb{E}[\mathbb{I}(\{\Delta_{S_t} \geq M\} \cap \mathcal{E}_{i,j}^{S_t}) \cdot \Delta_{S_t}] \\
&= \sum_{t=1}^{T} \left(\Pr[\{\Delta_{S_t} \geq M\} \cap \mathcal{E}_{i,j}^{S_t}] \cdot M + \int_M^{+\infty} \Pr[\{\Delta_{S_t} \geq x\} \cap \mathcal{E}_{i,j}^{S_t}] \, \mathrm{d}x \right) \\
&= \sum_{t=1}^{T} \Pr[\{\Delta_{S_t} \geq M\} \cap \mathcal{E}_{i,j}^{S_t}] \cdot M + \int_M^{+\infty} \sum_{t=1}^{T} \Pr[\{\Delta_{S_t} \geq x\} \cap \mathcal{E}_{i,j}^{S_t}] \, \mathrm{d}x \\
&= \sum_{t=1}^{T} \Pr[\{\Delta_{S_t} \geq M\} \cap \mathcal{E}_{i,j}^{S_t}] \cdot M + \int_M^{\max\{\Delta_{\max}^i, M\}} \sum_{t=1}^{T} \Pr[\{\Delta_{S_t} \geq x\} \cap \mathcal{E}_{i,j}^{S_t}] \, \mathrm{d}x \\
&\leq \ell_T(M, 2^{-j}) M + \int_M^{\max\{\Delta_{\max}^i, M\}} \ell_T(x, 2^{-j}) \, \mathrm{d}x.
\end{aligned}$$

□

F.3 Comparison between 1-norm and ∞-norm

In this paper, we give upper bounds of regret for CMAB-T problems that satisfy TPM bounded smoothness with 1-norm or with ∞-norm. We emphasis Theorem 1 and Theorem 5 do not imply each other. For clarity, we use a_1 and a_∞ in place of a in bounded smoothness function $f(x) = ax$. If a CMAB-T problem instance satisfies TPM bounded smoothness with 1-norm with $f(x) = a_1 x$, then it also satisfies TPM bounded smoothness with ∞-norm with $f(x) = a_\infty x$, where $a_\infty = K a_1$. Conversely, if a CMAB-T problem instance satisfies TPM bounded smoothness with ∞-norm with $f(x) = a_\infty x$, then it also satisfies TPM bounded smoothness with 1-norm with $f(x) = a_1 x$, where $a_1 = a_\infty$. For distribution-dependent upper bound, according to Theorems 1 and 5, we have $O(\frac{a_\infty^2 m \ln T}{\Delta})$ and $O(\frac{a_1^2 K m \ln T}{\Delta})$. For a problem instance that satisfies TPM bounded smoothness with 1-norm with $f(x) = a_1 x$, if we use the bound for ∞-norm, the result will be $O(\frac{a_1^2 K^2 m \ln T}{\Delta})$. For a problem instance that satisfies TPM bounded smoothness with ∞-norm with $f(x) = a_\infty x$, if we use the bound for 1-norm, the result will be $O(\frac{a_\infty^2 K m \ln T}{\Delta})$. Both give an additional K factor. It is similar for distribution-independent bound, which will have an additional $\sqrt{K}$ factor in both cases.