Batch-Size Independent Regret Bounds for the Combinatorial Multi-Armed Bandit Problem

Nadav Merlis   MERLIS@CAMPUS.TECHNION.AC.IL
Faculty of Electrical Engineering, Technion, Israel Institute of Technology.

Shie Mannor   SHIE@EE.TECHNION.AC.IL
Faculty of Electrical Engineering, Technion, Israel Institute of Technology.

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We consider the combinatorial multi-armed bandit (CMAB) problem, where the reward function is nonlinear. In this setting, the agent chooses a batch of arms on each round and receives feedback from each arm of the batch. The reward that the agent aims to maximize is a function of the selected arms and their expectations. In many applications, the reward function is highly nonlinear, and the performance of existing algorithms relies on a global Lipschitz constant to encapsulate the function's nonlinearity. This may lead to loose regret bounds, since by itself, a large gradient does not necessarily cause a large regret, but only in regions where the uncertainty in the reward's parameters is high. To overcome this problem, we introduce a new smoothness criterion, which we term *Gini-weighted smoothness*, that takes into account both the nonlinearity of the reward and concentration properties of the arms. We show that a linear dependence of the regret in the batch size in existing algorithms can be replaced by this smoothness parameter. This, in turn, leads to much tighter regret bounds when the smoothness parameter is batch-size independent. For example, in the probabilistic maximum coverage (PMC) problem, that has many applications, including influence maximization, diverse recommendations and more, we achieve dramatic improvements in the upper bounds. We also prove matching lower bounds for the PMC problem and show that our algorithm is tight, up to a logarithmic factor in the problem's parameters.

Keywords: Multi-Armed Bandits, Combinatorial Bandits, Probabilistic Maximum Coverage, Gini-Weighted Smoothness, Empirical Bernstein

1. Introduction

The multi-armed bandit (MAB) problem is one of the most elementary problems in decision making under uncertainty. Under this setting, the agent must choose an action (or arm) on each round, out of L possible actions. It then observes the chosen arm's reward, which is generated from some fixed unknown distribution, and aims to maximize the average cumulative reward (Robbins, 1952). Equivalently, the agent minimizes its expected cumulative regret, i.e., the difference between the best achievable reward and the agent's cumulative reward. This framework enables us to understand and control the trade-off between information gathering ('exploration') and reward maximization ('exploitation'), and many of the current reinforcement learning algorithms are based on exploration concepts that originate from MAB (Jaksch et al., 2010; Bellemare et al., 2016; Osband et al., 2013; Gopalan and Mannor, 2015).

One of the active research directions in MAB consists in extending the model to support more complicated feedbacks from the environment and more complex reward functions. An important

extension that follows this direction is the combinatorial multi-armed-bandit (CMAB) problem with semi-bandit feedback (Chen et al., 2016a). Instead of choosing a single arm, the agent selects a subset of the arms A_t (a 'batch'), and observes feedback $X_a(t)$ from each of the arms $a \in A_t$ ('semi bandit feedback'). The reward can be a general function of the expectations $\mu_a = \mathbb{E}[X_a(t)]$, with the linear function as the most common example: $r(A_t; \mu) = \sum_{a \in A_t} \mu_a$ for any batch of size $|A_t| \leq K$.

Another common case that falls under the CMAB framework, and will benefit from the results of this paper, is the bandit version of the probabilistic maximum coverage (PMC) problem. Under this setting, each arm is a random set that may contain some subset of M possible items, according to a fixed probability distribution. On each round, the agent chooses a batch of K sets and aims to maximize the number of items that appear in any of the sets (i.e., the size of the union of the sets). In the bandit setting, we assume that the probabilities that items appear in a set are unknown and aim to maximize the item coverage while concurrently learning the probabilities. Variants of the PMC bandit problem have many practical applications, such as influence maximization (Vaswani et al., 2015), ranked recommendations (Kveton et al., 2015a), wireless channel monitoring (Arora et al., 2011), online advertisement placement (Chen et al., 2016a) and more.

Although existing algorithms offer solutions to the CMAB framework under very general assumptions (Chen et al., 2016a,b), there are several issues that still pose a major challenge in the design and analysis of practical algorithms. Notably, most of the existing algorithms quantify the nonlinearity of the function using a global Lipschitz constant. However, large gradients do not necessarily translate to a large regret, but rather the combined influence of the gradient size and the local parameter uncertainty. More specifically, if there are regions in which tight parameter estimates can be derived, the regret will not be large even if the gradients are large. Conversely, in regions where the parameter uncertainty is large, smaller gradients can still cause a large regret. For example, consider the reward function $r(A; \mu) = \sum_{a \in A} \min\{c\mu_a, 1\}$, for $c \gg 1$ and $\mu_a \in [0, 1]$, and its translated version $r'(A; \mu) = \sum_{a \in A} \max\{\min\{c(\mu_a - \frac{1}{2}), 1\}, 0\}$. Despite the fact that both functions have the same Lipschitz constant, the regret in the first problem can be much smaller, since parameters are easier to estimate when they are close to the edge of their domain. Another similar example is the PMC bandit problem, in which the reward is approximately linear when the coverage probabilities are small, but declines exponentially when they are large. Thus, the naïve bound cannot capture the interaction between the gradient size and the parameter uncertainty, which results in loose regret bounds.

In this paper, we aim to utilize this principle for the CMAB framework. To this end, we introduce a new smoothness criterion called *Gini-weighted smoothness*. This criterion takes into account both the nonlinearity of the reward function and the concentration properties of bounded random variables around the edges of their domain. We then suggest an upper confidence bound (UCB) based strategy (Auer et al., 2002a), but replace the classical Hoeffding-type confidence intervals with ones that depend on the empirical variance of the arms, and are based on the Empirical Bernstein inequality (Audibert et al., 2009). We show that Bernstein-type bounds capture similar properties to the Gini-weighted smoothness, and thus allow us to derive tighter regret bounds. Notably, the linear dependence of the regret bound in the batch size is almost completely removed, except for a logarithmic factor, and the batch size only affects the regret through the Gini-smoothness parameter. In problems in which this parameter is batch-size independent, including the PMC bandit problem, our new bound is tighter by a factor of the batch size K. This is comparable to the best possible improvement due to independence assumption in the linear CMAB problem (Degenne and Perchet, 2016), but without any additional statistical assumptions.

2

Moreover, we demonstrate the tightness of our regret bounds by proving matching lower bounds for the PMC bandit problem, up to logarithmic factors in the batch size. To do so, we construct an instance of the PMC bandit problem that is equivalent to a classical MAB problem and then analyze the lower bounds of this problem. We also show that in contrast to the linear CMAB problem, the lower bounds do not change even if different sets are independent. To the best of our knowledge, our algorithm is the first to achieve tight regret bounds for the PMC bandit problem.

2. Related Work

The multi-armed bandits' literature is vast. We thus cover only some aspects of the area, and refer the reader to (Bubeck et al., 2012) and (Lattimore and Szepesvári, 2018) for a comprehensive survey. We employ the Optimism in the Face of Uncertainty (OFU) principle (Lai and Robbins, 1985), which is one of the most fundamental concepts in MAB, and can be found in many known MAB algorithms (e.g., Auer et al. 2002a; Garivier and Cappé 2011). While many algorithms rely on Hoeffding-type concentration bounds to derive an upper confidence bound (UCB) of an arm, a few previous works also apply Bernstein-type bounds and demonstrate superior performance, both in theory and in practice (Audibert et al., 2009; Mukherjee et al., 2018).

The general stochastic combinatorial multi-armed bandit framework was introduced in (Chen et al., 2013). They presented CUCB, an algorithm that uses UCB per arm (or 'base arm'), and then inputs the optimistic value of the arm into a maximization oracle. Many preceding works also fall under the CMAB framework, but mainly focus on a specific reward function (Caro and Gallien, 2007; Gai et al., 2010, 2012; Liu and Zhao, 2012), or work in the adversarial setting (Cesa-Bianchi and Lugosi, 2012). While most algorithms for the CMAB setting follow the OFU principle, a few employ Thompson Sampling (Thompson, 1933), e.g., (Wang and Chen, 2018; Hüyük and Tekin, 2018) for the semi-bandit feedback and (Gopalan et al., 2014) for the full-bandit feedback. In recent years, there have been extensive studies on deriving tighter bounds, but these works mostly address the linear CMAB problem (Kveton et al., 2014, 2015c; Combes et al., 2015b; Degenne and Perchet, 2016). More recently, tighter regret bounds were derived for this framework in (Wang and Chen, 2017), and also allow probabilistically triggered arms. Nevertheless, we show that in our setting, these bounds can be improved by a factor of the batch size K for many problems, e.g., the PMC bandit problem, and are comparable, up to logarithmic factors, otherwise.

Empirical Bernstein Inequality was first used for the CMAB problem in (Gisselbrecht et al., 2015) for linear reward functions, and was later used in (Perrault et al., 2018) for sequential search-and-stop problems. Both works focus on specific reward functions and utilize Bernstein inequality to get variance-dependent regret bounds. In contrast, we analyze general reward functions and exploit the relation between the confidence interval and the reward function to derive tighter regret bounds.

The PMC bandit problem is the bandit version of the maximum coverage problem (Hochbaum, 1996), a well studied subject in computer science with many variants and extensions. The bandit variant is closely related to the influence maximization problem (Vaswani et al., 2015; Carpentier and Valko, 2016; Wen et al., 2017), in which the agent chooses a set of nodes in a graph, that influence other nodes through random edges, and aims to maximize the number of influenced nodes. Another related setting is the cascading bandit problem (e.g., Kveton et al. 2015a,b; Combes et al. 2015a; Lattimore et al. 2018), in which a list of items is sequentially shown to a user until she finds one of them satisfactory. This is equivalent to a coverage problem with a single object, but only partial feedback - the user will not give any feedback about items that appear after the one she liked. In both

settings, the focus is very different than ours. In influence maximization, the focus is on the diffusion inside a graph and the graph structure, and in cascading bandits on the partial feedback and the list ordering. They are thus complementary to our framework and could benefit from our results.

3. Preliminaries

We work under the stochastic combinatorial-semi bandits framework, when the reward is the weighted sum of smooth monotonic functions. Assume that there are L arms ('base arms'), and let $\mathcal{A} \subset 2^{[L]}$ be the action set, i.e., the collection of possible batches (actions) from which the agent can choose, with $[L] = \{1, \ldots, L\}$. Also assume that the size of any batch $A \in \mathcal{A}$ is bounded by $|A| \leq K$. Denote the reward function of an action A with arm parameters p by $r(A; p)$ and assume that the reward is the weighted sum of M functions, $r(A; p) = \sum_{i=1}^{M} w_i r_i(A; p_i)$, for some fixed weights $\{w_i\}_{i=1}^{M}, w_i \geq 0$ and $p_i \in \mathbb{R}^L$. Without loss of generality, we also assume $r_i(A; x) \geq 0$.

The agent interacts with the environment as follows: on each round t, the agent chooses an action $A_t \in \mathcal{A}$. Then, for each arm $j \in A_t$, it observes feedback $X_{ij}(t) \in [0, 1]$ for any $i \in [M]$, with mean $\mathbb{E}[X_{ij}(t)] = p_{ij}$. For ease of notation, assume that $X_{ij}(t) = 0$ if $j \notin A_t$. We denote the empirical estimators of the parameters p_{ij} by $\hat{p}_{ij}(t) = \frac{1}{N_j(t)} \sum_{\tau=1}^{t} X_{ij}(\tau)$, where $N_j(t) = \sum_{\tau=1}^{t} \mathbb{1}\{j \in A_\tau\}$ is the number of times an arm j was chosen up to time t, and $\mathbb{1}\{\cdot\}$ is the indicator function. We also denote the empirical variance by $\hat{V}_{ij}(t) = \frac{1}{N_j(t)} \sum_{\tau=1}^{t} X_{ij}^2(\tau) - (\hat{p}_{ij}(t))^2$. The estimated parameters $\hat{p}_{ij}(t)$ are concentrated around their mean according to the Empirical Bernstein inequality:

Lemma 1 (Empirical Bernstein) *(Audibert et al., 2009) Let $X_1, \ldots, X_s$ be i.i.d random variables taking their values in $[0, 1]$, and let $\mu = \mathbb{E}[X_i]$ be their common expected value. Consider the empirical mean $\bar{X}_s$ and variance V_s defined respectively by*

$$\bar{X}_s = \frac{1}{s} \sum_{i=1}^{s} X_i \quad and \quad V_s = \frac{1}{s} \sum_{i=1}^{s} \left(X_i - \bar{X}_s\right)^2 = \frac{1}{s}\left(\sum_{i=1}^{s} X_i^2\right) - \bar{X}_s^2.$$

Then, for any $s \in \mathbb{N}$ and $x > 0$, it holds that $\Pr\left\{ |\bar{X}_t - \mu| \geq \sqrt{\frac{2V_t x}{t}} + \frac{3x}{t} \right\} \leq 3e^{-x}.$

We require the functions $r_i(A; p_i)$ to be monotonic Gini-smooth with smoothness parameters γ_∞, γ_g, which we define in the following:

Definition 2 *Let $f(A; x) : \mathcal{A} \times [0, 1]^L \to \mathbb{R}$ be a differentiable function in $x \in (0, 1)^L$ and continuous in $x \in [0, 1]^L$, for any $A \in \mathcal{A}$. The function $f(A; x)$ is said to be monotonic Gini-smooth, with smoothness parameters γ_∞ and γ_g, if:*

1. *For any $A \in \mathcal{A}$, the function is monotonically increasing with bounded gradient, i.e., for any $i \in A$ and $x \in (0, 1)^L$, $0 \leq \frac{\partial f(A; x)}{\partial x_i} \leq \gamma_\infty$. If $i \notin A$, then $\frac{\partial f(A; x)}{\partial x_i} = 0$ for all $x \in (0, 1)^L$.*

2. *For any $A \in \mathcal{A}$ and $x \in (0, 1)^L$, it holds that*

$$\sqrt{\sum_{i=1}^{L} x_i(1 - x_i)\left(\frac{\partial f(A; x)}{\partial x_i}\right)^2} \leq \gamma_g \ . \tag{1}$$

Throughout the paper, we refer this condition as the Gini-weighted smoothness[1] of $f(A; x)$.

---

1. The name is motivated by the similarity of the weights to the Gini impurity $\sum_i x_i(1 - x_i)$.

While the first condition is very intuitive, and is equivalent to the standard smoothness requirement (Wang and Chen, 2017), the second condition demands further explanation. Notably, the Gini-smoothness parameter is less sensitive to changes in the reward function when the parameters are close to the edges, i.e., close to 0 or 1. It may be observed that in these regions, the variances of the parameters are small, which implies that they are more concentrated around their mean. We will later show that this will allow us to mitigate the effect of large gradients on the regret. For simplicity, we assume that all of the functions $r_i(A; p_i)$ have the same smoothness parameters, but the extension is trivial. We note that if for all $i \in [M]$, the functions $r_i(A; p_i)$ are Gini-smooth, then $r(A; p)$ is also Gini-smooth. Nevertheless, explicitly decomposing the reward into a sum of monotonic Gini-smooth functions leads to a slightly tighter regret bound. We believe that this is due to a technical artefact, but nonetheless modify our analysis since many important cases fall under this scenario.

An important example that falls under our model is the Probabilistic Maximum Coverage (PMC) problem. In this setting, each arm is a random set that may contain some subset of M possible items. The agent's goal is to choose a batch of sets such that as many items appear in the sets (i.e., the union of the sets is maximized). Formally, the functions $r_i(A; p) = 1 - \prod_{j \in A}(1 - p_{ij})$ are the probabilities that an item $i \in [M]$ was covered and $r(A; p)$ is the (weighted) expected number of covered items. The smoothness constants for this problem are $\gamma_\infty = 1$ and $\gamma_g = \frac{1}{\sqrt{e}}$, independently of the batch size. Another example is the logistic function $r_i(A; p) = \frac{1}{1 + Ce^{-\sum_{j \in A} p_{ij}}}$, for which the smoothness parameters are equal to $\gamma_\infty = \frac{1}{4}$ and $\gamma_g = \frac{1}{4}\sqrt{1 + \log C}$, which are batch size independent, provided that C does not depend on K. In the linear case, $r_i(A; p) = r(A; p) = \sum_{j \in A} p_j$ and the smoothness parameters are $\gamma_\infty = 1$ and $\gamma_g = \sqrt{K/4}$. In general, we can always bound $\gamma_g \leq \frac{1}{2}\sqrt{K}\gamma_\infty$, and we will later see that in this case, our bounds are comparable to existing results that only rely on γ_∞.

Similarly to previous work, the performance of an agent is measured according to its regret, i.e., the difference between the reward of an oracle, which acts optimally according to the statistics of the arms, and the cumulative reward of the agent. However, in many problems, it is impractical to achieve the optimal solution even when the problem's parameters are known. Thus, it is more sensible to compare the performance of an algorithm to the best approximate solution that an oracle can achieve. Denote the optimal action by $A^* = \arg\max_{A \in \mathcal{A}}\{r(A; p)\}$ and its value by $r_{\max}$. An oracle is called an (α, β) *approximation oracle* if for every parameter set p, with probability β, it outputs a solution whose value is at least $\alpha r_{\max}$. We define the *expected approximation regret* of an algorithm as:

$$R(T) = \alpha\beta r_{\max}T - \sum_{t=1}^{T} \mathbb{E}[r(A_t; p)] \ , \tag{2}$$

where the expectation is over the randomness of the environment and the agent's actions, through the oracle. As noted by (Chen et al., 2016a), in the linear case, and sometimes when arms are independent, the reward function equals to the expectation of the empirical reward, i.e., $r(A_t; \mu) = \mathbb{E}[r(A_t; X)]$ but unfortunately, it does not necessarily occur when arms are arbitrarily correlated.

We end this section with some notations. Let $\bar{M} = \sum_{i=1}^{M} w_i$. Denote the suboptimality gap of a batch A_t by $\Delta_{A_t} = \alpha r_{\max} - r(A_t; p)$. The minimal gap of a base arm j is the smallest positive gap of a batch A that contains arm j, namely, $\Delta_{j,\min} = \min_{A \in \mathcal{A}, j \in A, \Delta_A > 0} \Delta_A$ and the maximal gap is $\Delta_{j,\max} = \max_{A \in \mathcal{A}, j \in A, \Delta_A > 0} \Delta_A$. Note that for all $A \in \mathcal{A}$, $\Delta_A \leq \max_j \Delta_{j,\max} \triangleq \Delta_{\max}$. We denote by $D_{\mathrm{KL}}(X, Y)$ the Kullback-Leibler (KL) divergence between two random variables X, Y.

Also denote by $\mathrm{kl}(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1-p}{1-q}$, the KL divergence between two Bernoulli random variables of means p and q.

4. Algorithm

Algorithm 1 BC-UCB (**B**ernstein **C**ombinatorial - **U**pper **C**onfidence **B**ound)

1: Initialize arm counts $N_j(0) = 0$
2: **for** $t = 1, \ldots, T$ **do**
3:     **if** $\exists j \in [L] : N_j(t - 1) = 0$ **then**
4:         Choose batch A_t such that $j \in A_t$
5:     **else**
6:         $q_{ij}(t) = \min\left\{1, \hat{p}_{ij}(t - 1) + \sqrt{\frac{6\hat{V}_{ij}(t-1)\log t}{N_j(t-1)}} + \frac{9\log t}{N_j(t-1)}\right\}$
7:         $A_t = \mathrm{Oracle}(q)$
8:     **end if**
9:     Play A_t and observe $X_{ij}(t)$ for all $j \in A_t$ and $i \in [M]$
10:    Update $N_j(t), \hat{p}_{ij}(t)$ and $\hat{V}_{ij}(t)$ for all $j \in A_t$ and $i \in [M]$
11: **end for**

We suggest a combinatorial UCB-type algorithm with Bernstein based confidence interval, which we call BC-UCB (**B**ernstein **C**ombinatorial - **U**pper **C**onfidence **B**ound). The UCB index is defined as

$$q_{ij}(t) = \min\left\{1, \hat{p}_{ij}(t - 1) + \sqrt{\frac{6\hat{V}_{ij}(t-1)\log t}{N_j(t-1)}} + \frac{9\log t}{N_j(t-1)}\right\} . \tag{3}$$

A pseudocode can be found in Algorithm 1. On the first rounds, the agent initially samples batches to make sure that each base arm is sampled at least once. Afterwards, we calculate the UCB index q, and an approximation oracle chooses an action A_t such that $r(A_t; q) \geq \alpha r(A^*(q); q)$ with probability β. The agent then plays this action and observes feedback for any arm $j \in A_t$. It finally updates the empirical probabilities and variances and continues to the next round.

An example for an approximation oracle that can be used when the reward function is monotonic and submodular [2] is the greedy oracle, which enjoys an approximation factor of $\alpha = 1 - \frac{1}{e}$ with probability $\beta = 1$ (Nemhauser et al., 1978). First, the oracle initializes $A_t(0) = \emptyset$, and then selects K items sequentially in a greedy manner, i.e. $j_n \in \arg\max_{j \notin A_t(n-1)}\{r(A_t(n - 1) \cup \{j\}; q)\}$ and $A_t(n) = A_t(n - 1) \cup \{j_n\}$, with $A_t = A_t(K)$.

The first term of the confidence interval resembles the standard UCB term of $\sqrt{\frac{3\log t}{2N_j(t-1)}}$, and is actually always smaller, since the empirical variance of variables in $[0, 1]$ is always bounded by $\frac{1}{4}$.[3] The second term does not appear in the standard UCB bound, and can slightly affect the regret, since suboptimal arms will be asymptotically sampled $O(\log t)$ times, so the two terms are comparable. Nevertheless, the first term is still dominant, and if the variance of an arm is drastically lower than $\frac{1}{4}$,

---

2. Let Ω be a finite set. A set function $f : 2^{\Omega} \to \mathbb{R}$ is called submodular if $\forall S, T \in \Omega, f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$.
3. $V = \frac{1}{n}\sum_i X_i^2 - \left(\frac{1}{n}\sum_i X_i\right)^2 \leq \frac{1}{n}\sum_i X_i - \left(\frac{1}{n}\sum_i X_i\right)^2 = \left(\frac{1}{n}\sum_i X_i\right)\left(1 - \frac{1}{n}\sum_i X_i\right) \leq \frac{1}{4}$.

the confidence bound is significantly tighter. This can happen, for example, when the arm's mean is close to 0 or 1.

We take advantage of this property through the smoothness parameter γ_g, that only takes into account the sensitivity of the function to parameter changes when the parameters are far away from 0 or 1. This will allow us to derive a drastically tighter regret bound, in comparison to existing algorithms, when $\gamma_g = o(\sqrt{K}\gamma_\infty)$, as we establish in the following theorem:

Theorem 3 *Let $r_i(A; p_i)$ be monotonic Gini-smooth reward functions with smoothnsess parameters γ_∞ and γ_g, and let $r(A; p) = \sum_{i=1}^M w_i r_i(A; p_i)$ be the reward function. For any $T \geq 1$, the expected approximation regret of BC-UCB with (α,β)-approximation oracle is bounded by*

$$R(T) \leq \left[8640\gamma_g^2\bar{M}^2 \sum_{j=1}^L \frac{1}{\Delta_{j,\min}} + 340\gamma_\infty\bar{M}\sum_{j=1}^L \left(1 + \log\frac{\Delta_{j,\max}}{\Delta_{j,\min}}\right) \right] \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T$$
$$+ L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right). \tag{4}$$

We can also exploit the problem dependent regret bound to derive a problem independent bound, that is, a bound that holds for any gaps $\Delta_{j,\min}$:

Corollary 4 *Let $r_i(A; p_i)$ be monotonic Gini-smooth reward functions with smoothnsess parameters γ_∞ and γ_g, and let $r(A; p) = \sum_{i=1}^M w_i r_i(A; p_i)$ be the reward function. For any $T \geq 1$, the expected approximation regret of BC-UCB with (α,β)-approximation oracle can be bounded by*

$$R(T) \leq 2\sqrt{8640}\gamma_g\bar{M}\left\lceil \frac{\log K}{1.61} \right\rceil \sqrt{LT\log T} + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right)$$
$$+ 340\gamma_\infty\bar{M}L\left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T \left(2 + \log\frac{\Delta_{\max}T}{340\gamma_\infty\bar{M}L\left\lceil \frac{\log K}{1.61}\right\rceil^2 \log T}\right). \tag{5}$$

The proof of Theorem 3 is presented in the following section, along with a proof sketch for Corollary 4. The full proof of the corollary can be found in Appendix E.

We start by noting that we could avoid decomposing the reward into sum of M functions, but the regret bound in this case is slightly looser - the $\bar{M}^2$ factor in (4) is replaced by the larger factor $M\sum_i w_i^2$, and the $\log K$ factor is replaced by $\log KM$. We believe that the logarithmic factor $\log K$ is due to a technical artefact, but leave its removal for future work. We also remark that the second term of the problem dependent regret bounds is negligible for small gaps and can always be bounded using the identity $1 + \log x \leq x$, which yields a regret of $O\left((\bar{M}^2\gamma_g^2 + \bar{M}\gamma_\infty\Delta_{\max})\left\lceil \frac{\log K}{1.61} \right\rceil^2 \sum_{j=1}^L \frac{\log T}{\Delta_{j,\min}}\right)$.

To the best of our knowledge, the closest bound to ours appears in (Wang and Chen, 2017). From their perspective, our bandit problem has ML base arms and a batch size of MK with gaps $\Delta_{ij,\min} = \Delta_{j,\min}$. Their L_1 smoothness parameter equals $B = \gamma_\infty \max_i w_i$. Substituting into their bound yields a regret of $O\left(\sum_{i,j} \frac{(\gamma_\infty \max_i w_i)^2 MK \log T}{\Delta_{ij,\min}}\right) = O\left(\gamma_\infty^2 K \sum_j \frac{(\max_i w_i)^2 M^2 \log T}{\Delta_{j,\min}}\right)$. Since $\gamma_g \leq \frac{1}{2}\sqrt{K}\gamma_\infty$, $\Delta_{\max} \leq \bar{M}K\gamma_\infty$ and $\bar{M} \leq M \max_i w_i$, our regret is tighter when γ_g is not trivial (that is, $\gamma_g = o(\sqrt{K}\gamma_\infty)$), up to a factor of $\log^2 K$.

Alternatively to our approach, it is possible to analyze BC-UCB only based on the γ_∞ smoothness. In this case, the analysis will be very similar to that of Kveton et al. (2015c). This will yield a dominant term that does not depend on the logarithmic factor $\log K$, and declines with the variance of the arms. However, it can lead to dramatically worse bounds when γ_g is small, so we decided not to pursue this path. Nonetheless, this approach is still worth mentioning when comparing to other algorithms, since when its bounds are combined with ours, we can conclude that the regret of BC-UCB is always tighter than the regret obtained in (Wang and Chen, 2017).

On a final note, we return to the PMC bandit problem. In this case, γ_g and γ_∞ are $O(1)$ and $\Delta_{\max}$ is $O(\bar{M})$, so the regret is $O\left(\bar{M}^2 \log^2 K \sum_{j=1}^L \frac{\log T}{\Delta_{j,\min}}\right)$, which is tighter by a factor of K from existing results (Wang and Chen, 2017). We will later show that this bound is tight, up to logarithmic factors in K.

5. Proving the Regret Upper Bounds

We start the proof by simplifying the first term of the UCB index. To this end, recall Bernstein's inequality:

Lemma 5 (Bernstein's Inequality) *Let $\{X_i\}_{i=1}^n$ be independent random variables in $[0,1]$ with mean $\mathbb{E}[X_i] = p$ and variance $\mathrm{Var}\{X\}$. Then, with probability $1 - \delta$:*

$$\frac{1}{n} \sum_{i=1}^n X_i \le p + \frac{2 \log 1/\delta}{3n} + \sqrt{\frac{2\mathrm{Var}\{X\} \log 1/\delta}{n}} \ . \tag{6}$$

Next, let $V = \frac{1}{n} \sum_{i=1}^n \left(X_i - \left(\frac{1}{n} \sum_{i=1}^n X_i\right)\right)^2$ be the empirical variance of independent random variables $X_i \in [0,1]$ with mean p, and note that $V = \frac{1}{n} \sum_{i=1}^n (X_i - p)^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i - p\right)^2 \le \frac{1}{n} \sum_{i=1}^n (X_i - p)^2$. We can thus define the independent random variables $Y_i = (X_i - p)^2$ and bound $\frac{1}{n} \sum_{i=1}^n Y_i$ instead. It is clear that $0 \le Y_i \le 1$, and their expectations can also be bounded by

$$\mathbb{E}[Y_i] = \mathrm{Var}\{X_i\} = \mathbb{E}X_i^2 - p^2 \le 1 \cdot \mathbb{E}X_i - p^2 = p(1-p) \ .$$

The variance of Y_i can be similarly bounded by $\mathrm{Var}\{Y_i\} \le \mathbb{E}[Y_i^2] \le 1 \cdot \mathbb{E}[Y_i] \le p(1-p)$. Applying Bernstein's Inequality (6) on Y_i can now give a high probability bound on the empirical variance: with probability $1 - \delta$,

$$V \le \frac{1}{n} \sum_{i=1}^n Y_i \le \mathbb{E}[Y_i] + \frac{2 \log 1/\delta}{3n} + \sqrt{\frac{2\mathrm{Var}\{Y_i\} \log 1/\delta}{n}}$$

$$\le p(1-p) + \frac{2 \log 1/\delta}{3n} + \sqrt{\frac{2p(1-p) \log 1/\delta}{n}}$$

$$\overset{(*)}{\le} p(1-p) + \frac{2 \log 1/\delta}{3n} + p(1-p) + \frac{\log 1/\delta}{2n}$$

$$= 2p(1-p) + \frac{7 \log 1/\delta}{6n} \ , \tag{7}$$

where $(*)$ utilizes the relation $ab \le \frac{1}{2}(a^2 + b^2)$ with $a = \sqrt{2p(1-p)}$ and $b = \sqrt{\frac{\log 1/\delta}{n}}$. An important (informal) conclusion of inequality (7), is that if the event under which the inequality holds for $\hat{V}_{ij}$ does occur, then the confidence interval around $\hat{p}_{ij}(t-1)$ can be bounded by:

$$\sqrt{\frac{6\hat{V}_{ij}(t-1)\log t}{N_j(t-1)}} + \frac{9\log t}{N_j(t-1)} \leq \sqrt{\frac{12p_{ij}(1-p_{ij})\log t}{N_j(t-1)} + \frac{7\log\frac{1}{\delta}\log t}{N_j^2(t-1)}} + \frac{9\log t}{N_j(t-1)}$$

$$\leq \sqrt{\frac{12p_{ij}(1-p_{ij})\log t}{N_j(t-1)}} + \frac{\sqrt{7\log\frac{1}{\delta}\log t}}{N_j(t-1)} + \frac{9\log t}{N_j(t-1)} \quad . \quad (8)$$

For $\delta \approx t^{-3}$, the confidence interval is of the form $u_i\sqrt{p_i(1-p_i)} + v_i$, for $u_i = O\left(\sqrt{\frac{\log t}{N_j(t-1)}}\right)$ and $v_i = O\left(\frac{\log t}{N_j(t-1)}\right)$. We should therefore analyze how this kind of parameter perturbations affect the reward function. To do so, we take advantage of the Gini-weighted smoothness, as stated in the following lemma (see Appendix A for proof):

Lemma 6 *Let $f(A;x)$ be a monotonic Gini-smooth function, with smoothness parameters γ_∞ and γ_g. Also let $u, v \in \mathbb{R}^L$ be some constant vectors such that $u_i \geq 0$ and $v_i \geq 0$.*

For any $x, \epsilon \in [0,1]^L$ such that $\epsilon_i \leq \min\left\{u_i\sqrt{x_i(1-x_i)} + v_i, 1 - x_i\right\}$, the sensitivity of $f(A;x)$ to parameter change can be bounded by

$$f(A; x+\epsilon) - f(A; x) \leq 3\sqrt{2}\gamma_g\sqrt{\sum_{i\in A}u_i^2} + \gamma_\infty\sum_{i\in A}v_i \quad . \quad (9)$$

Next, define the low probability events under which some of the variables are not concentrated in their confidence intervals:

$$\mathbb{H}_t^p = \left\{\exists i,j : |\hat{p}_{ij}(t-1) - p_{ij}| > \sqrt{\frac{6\hat{V}_{ij}(t-1)\log t}{N_j(t-1)}} + \frac{9\log t}{N_j(t-1)}\right\} \quad (10)$$

$$\mathbb{H}_t^V = \left\{\exists i,j : \hat{V}_{ij}(t-1) > 2p(1-p) + 3.5\frac{\log t}{N_j(t-1)}\right\} \quad (11)$$

Also denote $\mathbb{H}_t = \mathbb{H}_t^V \cup \mathbb{H}_t^p$. Intuitively, even though the regret may be large under $\mathbb{H}_t$, the event cannot occur many times, and we can therefore analyze the regret under the assumption that $\mathbb{H}_t$ does not occur. We can then bound the regret similarly to inequality (8), combined with Lemma 6. Formally, we decompose the regret as follows:

Lemma 7 *Let $r_i(A;p)$ be Gini-smooth functions with parameters γ_g, γ_∞, and define*

$$c_t(A_t) = c_1\sqrt{\sum_{j\in A_t}\frac{\log t}{N_j(t-1)}} + c_2\sum_{j\in A_t}\frac{\log t}{N_j(t-1)} \quad (12)$$

for $c_1 = 12\sqrt{6}\gamma_g$ and $c_2 = 34\gamma_\infty$. The regret of Algorithm 1, when used with (α, β) approximation oracle, can be bounded by

$$R(T) \leq \mathbb{E}\left[\sum_{t=L+1}^{T}\Delta_{A_t}\mathbb{1}\left\{\Delta_{A_t} \leq \bar{M}c_t(A_t)\right\}\right] + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right). \quad (13)$$

9

The proof is in Appendix B. It is interesting to observe that $c_t(A_t)$ is very similar in its form to the confidence interval for the linear combinatorial problem when parameters are independent (Combes et al., 2015b). We have achieved this form of confidence without any independence assumptions, only on the basis of the properties of the reward function. We can therefore adapt the proofs of (Kveton et al., 2015c; Degenne and Perchet, 2016) to derive a problem dependent regret bound. Since we are interested in bounding the regret of the first term, and due to the initial sampling stage, we assume from this point onward that all of the arms were sampled at least once.

Define $\{a_k\}_{k=0}^\infty$ and $\{b_k\}_{k=0}^\infty$, two positive decreasing sequences that converge to 0 and will be determined later, with $b_0 = 1$. Also define the set $S_t^k = \left\{ j \in A_t : N_j(t-1) \le a_k \frac{g(K, \Delta_{A_t}) \log t}{\Delta_{A_t}^2} \right\}$ for some function $g(K, \Delta_{A_t})$ that will also be determined later, with $S_t^0 = A_t$. Intuitively, S_t^k is the set of arms that were chosen on round t and were not sampled enough times. Denote the events in which S_t^k contains at least $K b_k$ elements, but S_t^n contain less than $K b_n$ for any $n < k$, by $\mathbb{G}_t^k = \left\{ \{|S_t^k| \ge K b_k\} \cap \{\forall n < k, |S_t^n| < K b_n\} \right\}$, and let $\mathbb{G}_t = \cup_{k=1}^\infty \mathbb{G}_t^k$. We show that when $\Delta_{A_t} \le \bar{M} c_t(A_t)$, then $\mathbb{G}_t$ must occur, for the appropriate $g(K, \Delta_{A_t})$. To do so, we first cite a variant of Lemmas 7 and 8 of (Degenne and Perchet, 2016), with $e = 1$, $\Gamma^{(ii)} = 1$ and $f(t) = \log t / 8$:

Lemma 8 *Define* $\ell = \left(\sum_{k=1}^{k_0} \frac{b_{k-1} - b_k}{a_k} + \frac{b_{k_0}}{a_{k_0}} \right)$, *and let* k_0 *be the smallest index such that* $b_{k_0} \le 1/K$. *Then* $\mathbb{G}_t = \cup_{k=1}^{k_0} \mathbb{G}_t^k$. *Also, under* $\bar{\mathbb{G}}_t$, *the following inequality holds:*

$$\sum_{j \in A_t} \frac{\log t}{N_j(t-1)} < \frac{K \Delta_{A_t}^2 \ell}{g(K, \Delta_{A_t})} \ . \tag{14}$$

Using this lemma, we can now prove that $\mathbb{G}_t$ must occur (see Appendix C for proof):

Lemma 9 *If* $g(K, \Delta_{A_t}) = \left(864 \gamma_g^2 + 68 \frac{\gamma_\infty \Delta_{A_t}}{\bar{M}} \right) \bar{M}^2 K \ell$, *and if* $\Delta_{A_t} \le \bar{M} c_t(A_t)$, *then* $\mathbb{G}_t$ *occurs.*

A direct result of this lemma is that if $\Delta_{A_t} \le \bar{M} c_t(A_t)$, at least one event $\{\mathbb{G}_t^k\}_{k=1}^{k_0}$ occurs. This allows us to further decompose the first term of (13), and achieve the final result of Theorem 3:

Lemma 10 *The regret from the event* $\Delta_{A_t} \le \bar{M} c_t(A_t)$ *can be bounded by*

$$\sum_{t=L+1}^T \Delta_{A_t} \mathbb{1}\{\Delta_{A_t} \le \bar{M} c_t(A_t)\}$$

$$\le \left[1728 \gamma_g^2 \bar{M}^2 \sum_{j=1}^L \frac{1}{\Delta_{j,\min}} + 68 \gamma_\infty \bar{M} \sum_{j=1}^L \left(1 + \log \frac{\Delta_{j,\max}}{\Delta_{j,\min}} \right) \right] \left(\sum_{k=1}^{k_0} \frac{a_k}{b_k} \right) \ell \log T \ , \tag{15}$$

and if $a_k = b_k = 0.2^k$,

$$\sum_{t=L+1}^T \Delta_{A_t} \mathbb{1}\{\Delta_{A_t} \le \bar{M} c_t(A_t)\}$$

$$\le \left[8640 \gamma_g^2 \bar{M}^2 \sum_{j=1}^L \frac{1}{\Delta_{j,\min}} + 340 \gamma_\infty \bar{M} \sum_{j=1}^L \left(1 + \log \frac{\Delta_{j,\max}}{\Delta_{j,\min}} \right) \right] \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T \ . \tag{16}$$

10

The proof is in Appendix D. Substituting (16) into (13) concludes the proof of Theorem 3. ∎

The proof for the problem independent upper bound of Corollary 4 is a direct result of Lemma 10. Specifically, the bound can be achieved by decomposing the regret according to Lemma 7, and then dividing the regret into large gaps ($\Delta_{A_t} \geq \Delta$) and small gaps ($\Delta_{A_t} \leq \Delta$), according to some fixed threshold Δ. Large gaps are bounded according to Lemma 10 with $\Delta_{j,\min} \geq \Delta$, and small gaps are bounded trivially by ΔT. The final bound is achieved by optimizing the threshold Δ. The full proof can be found in Appendix E.

6. Lower Bounds

Although our algorithm enjoys improved upper bounds in comparison to CUCB, it is still interesting to see whether our results are tight in problems where previous bounds are loose. To demonstrate the tightness of our algorithm, we present an instance of the PMC bandit problem, on which our results are tight up to logarithmic factors. We assume throughout the rest of this section that the maximization oracle can output the optimal batch, i.e. has an approximation factor of $\alpha = 1$, with probability $\beta = 1$. This assumption allows us to focus on the difficulty of the problem due to parameter uncertainty and the semi-bandit feedback. We formally state the results in the following proposition:

Proposition 11 *There exist an instance of the PMC bandit problem with minimal gap Δ such that the expected regret of any consistent algorithm[4] is bounded by*

$$\liminf_{t \to \infty} \frac{R(t)}{\log t} \geq \frac{\bar{M}^2(L - K)}{4\Delta} \quad . \tag{17}$$

Moreover, for any $T > 0$ and $L > K$, there exist an instance of the PMC bandit problem such that the expected regret of any algorithm is bounded by

$$R(T) \geq \frac{1}{20} \bar{M} \min\left\{ \sqrt{(L - K + 1)T}, T \right\} \quad . \tag{18}$$

Proof

Consider the following PMC bandit problem: fix the first $K - 1$ arms to be empty sets, that is, $X_{ij} = 0$ for any $(i, j) \in [M] \times [K - 1]$, which also implies $p_{ij} = 0$. For the rest of the arms $j \in \{K, \ldots, L\}$, we force all of the items to be identically distributed, i.e., $X_{ij} = X_j$ and $p_{ij} = p_j$. We also fix the action set to be $\mathcal{A} = \{A_j\}_{j=1}^{L-K+1}$, where $A_j = \{1, \ldots, K - 1, j + K - 1\}$ contains all of the $K - 1$ arms with zero reward plus an additional arm $j + K - 1$. The expected reward when choosing an action A_j is thus $\bar{M}p_j \triangleq \mu_j$, and the problem is equivalent to a $(L - K + 1)$-armed bandit problem with arm distribution $Y_j = \sum_i w_i X_{j+K-1} = \bar{M} X_{j+K-1}$. In order to prove the problem dependent regret bound, let $p_j = \frac{1}{2} - \epsilon, \forall j \in \{K, \ldots, L - 1\}$ and $p_L = \frac{1}{2}$, or equivalently, $\mu_j = \frac{\bar{M}}{2} - \bar{M}\epsilon, \forall j \in [L - K]$ and $\mu_{L-K+1} = \frac{\bar{M}}{2}$. The gaps of the problem are thus $\Delta_j = \bar{M}\epsilon \triangleq \Delta$ for any arm $j \in [L - K]$. For any consistent MAB algorithm, the expected regret of the algorithm can be lower bounded by (Lai and Robbins, 1985):

$$\liminf_{t \to \infty} \frac{R(t)}{\log t} \geq \sum_{j=1}^{L-K} \frac{\Delta_j}{D_{\mathrm{KL}}(Y_j, Y_{L-K+1})} \quad . \tag{19}$$

---

4. An algorithm is called consistent if for any problem and any $\alpha > 0$, the regret of the algorithm is $R(t) = o(t^\alpha)$ as $t \to \infty$.

The KL divergence can be directly bounded by

$$D_{\mathrm{KL}}(Y_j, Y_{L-K+1}) = \mathrm{kl}(p_j, p_{L-K+1}) = \mathrm{kl}\left(\frac{1}{2} - \epsilon, \frac{1}{2}\right) \overset{(*)}{\leq} 4\epsilon^2 = 4\frac{\Delta^2}{\bar{M}^2} \ ,$$

where $(*)$ is due to the relation $\mathrm{kl}(p, q) \leq \frac{(p-q)^2}{q(1-q)}$ (Csiszár and Talata, 2006). Substituting back into (19) yields the first part of the proposition:

$$\lim_{t \to \infty} \inf \frac{R(t)}{\log t} \geq \sum_{j=1}^{L-K} \frac{\Delta_j}{D_{\mathrm{KL}}(Y_j, Y_{L-K+1})} \geq \sum_{j=1}^{L-K} \frac{\Delta}{4\Delta^2/\bar{M}^2} = \frac{\bar{M}^2(L-K)}{4\Delta} \ .$$

Due to the fact that all of the items have the same distribution, our problem is equivalent to an $(L - K + 1)$ MAB problem with arms in $[0, 1]$ scaled by a factor of $\bar{M}$. Thus, the problem independent lower bound from (Auer et al., 2002b) can also be applied to this problem, and is scaled by the same factor $\bar{M}$:

$$R(T) \geq \frac{1}{20}\bar{M} \min\left\{ \sqrt{(L-K+1)T}, T \right\} \ .$$

∎

We remark that throughout the proof, we assumed nothing about the correlation between different arms, and thus the bound cannot be improved by assuming this kind of independence. Nevertheless, if items are assumed to be independent, the lower bounds can be drastically improved. We will not tackle the problem of independent items in this paper, but leave it to future work.

7. Summary

In this work, we introduced BC-UCB, a CMAB algorithm that utilizes Bernstein-type confidence intervals. We defined a new smoothness criterion called Gini-weighted smoothness, and showed that it allows us to derive tighter regret bounds in many interesting problems. We also presented matching lower bounds for such a problem, and thus demonstrated the tightness of our algorithm.

We believe that our concepts can be applied to derive tighter bounds in many interesting settings. Specifically, our analysis includes the PMC bandit problem, that has a central place in the areas of ranked recommendations and influence maximization. We also believe that our results could be extended to the frameworks of cascading bandits and probabilistically triggered arms, but leave this for future work.

Another possible direction involves analyzing specific arm distributions - in our framework, we assumed nothing about the arms' distribution except for its domain, and thus could only take into account very weak concentration properties, and specifically the concentration properties around the edges of the domain. If additional information about the distribution of the arms is present, it should be possible to leverage such information to design more sophisticated smoothness criteria. Such criteria could take into account tighter concentration properties of the arms' distribution, and thus lead to tighter regret bounds.

Finally, we remark that the lower bounds were possible to derive only since we have required the algorithm to support any arbitrary choice of action set $\mathcal{A}$. For the PMC problem, previous work shows that when $\mathcal{A}$ contains any subset of fixed size K, the regret bounds can be significantly improved (Kveton et al., 2015a). It is interesting to see if our technique can be used in this setting to extend these results and derive tighter bounds for any Gini-smooth function.

Acknowledgments

The authors thank Asaf Cassel and Esther Derman for their helpful comments on the manuscript.

References

Pallavi Arora, Csaba Szepesvári, and Rong Zheng. *Sequential learning for optimal monitoring of multi-channel wireless networks*. IEEE, 2011.

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

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016.

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

Felipe Caro and Jérémie Gallien. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

Alexandra Carpentier and Michal Valko. Revealing graph bandits for maximizing local influence. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159, 2013.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1): 1746–1778, 2016a.

Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In *Advances in Neural Information Processing Systems*, pages 1659–1667, 2016b.

Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. Learning to rank: Regret lower bounds and efficient algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):231–244, 2015a.

Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124, 2015b.

Imre Csiszár and Zsolt Talata. Context tree estimation for not necessarily finite memory processes, via bic and mdl. *IEEE Transactions on Information theory*, 52(3):1007–1016, 2006.

Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, pages 2972–2980, 2016.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*, pages 1–9. IEEE, 2010.

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 (TON)*, 20(5):1466–1478, 2012.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

Thibault Gisselbrecht, Ludovic Denoyer, Patrick Gallinari, and Sylvain Lamprier. Whichstreams: A dynamic approach for focused data capture from large social media. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

Aditya Gopalan and Shie Mannor. Thompson sampling for learning parameterized markov decision processes. In *Conference on Learning Theory*, pages 861–898, 2015.

Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pages 100–108, 2014.

Dorit S Hochbaum. *Approximation algorithms for NP-hard problems*. PWS Publishing Co., 1996.

Alihan Hüyük and Cem Tekin. Thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms. *arXiv preprint arXiv:1809.02707*, 2018.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 420–429. AUAI Press, 2014.

Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, 2015a.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. In *Advances in Neural Information Processing Systems*, pages 1450–1458, 2015b.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543, 2015c.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Tor Lattimore, Branislav Kveton, Shuai Li, and Csaba Szepesvari. Toprank: A practical algorithm for online stochastic ranking. In *Advances in Neural Information Processing Systems*, pages 3949–3958, 2018.

Keqin Liu and Qing Zhao. Adaptive shortest-path routing under unknown and stochastically varying link states. In *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012 10th International Symposium on*, pages 232–237. IEEE, 2012.

Subhojyoti Mukherjee, KP Naveen, Nandan Sudarsanam, and Balaraman Ravindran. Efficient-ucbv: An almost optimal algorithm using variance estimates. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Pierre Perrault, Vianney Perchet, and Michal Valko. Finding the bandit in a graph: Sequential search-and-stop. *arXiv preprint arXiv:1806.02282*, 2018.

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

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.

Sharan Vaswani, Laks Lakshmanan, Mark Schmidt, et al. Influence maximization with bandits. *arXiv preprint arXiv:1503.00024*, 2015.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, pages 1161–1171, 2017.

Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5101–5109, 2018.

Zheng Wen, Branislav Kveton, Michal Valko, and Sharan Vaswani. Online influence maximization under independent cascade model with semi-bandit feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3026–3036. Curran Associates Inc., 2017.

Appendix A. Proof of Lemma 6

Lemma 6 *Let $f(A; x)$ be a monotonic Gini-smooth function, with smoothness parameters γ_∞ and γ_g. Also let $u, v \in \mathbb{R}^L$ be some constant vectors such that $u_i \geq 0$ and $v_i \geq 0$.*

For any $x, \epsilon \in [0, 1]^L$ such that $\epsilon_i \leq \min\left\{u_i\sqrt{x_i(1 - x_i)} + v_i, 1 - x_i\right\}$, the sensitivity of $f(A; x)$ to parameter change can be bounded by

$$f(A; x + \epsilon) - f(A; x) \leq 3\sqrt{2}\gamma_g\sqrt{\sum_{i \in A} u_i^2} + \gamma_\infty \sum_{i \in A} v_i \ . \tag{9}$$

Proof

First, we define the functions

$$g(z) = \int_0^z \frac{dy}{\sqrt{y(1 - y)}} \quad , \quad h(z) = \int_0^z \frac{dy}{\min\{\sqrt{y}, \sqrt{1 - y}\}} \ . \tag{20}$$

We note that $g(z)$ is well defined for $z \in [0, 1]$, since the function $1/\sqrt{y}$ is integrable near $y = 0$ and $1/\sqrt{1 - y}$ is continuous, so the product $\frac{1}{\sqrt{y(1-y)}}$ is integrable near $z = 0$. Symmetrically, the function is also integrable near $y = 1$. $h(z)$ can be explicitly written as

$$h(z) = \begin{cases} 2\sqrt{z} & , z \leq \frac{1}{2} \\ 2\sqrt{2} - 2\sqrt{1 - z} & , z \geq \frac{1}{2} \end{cases} \tag{21}$$

The two functions are closely related: observe that $h'(z) \leq g'(z) \leq \sqrt{2}h'(z)$ with $g(0) = h(0) = 0$, and thus $h(z) \leq g(z) \leq \sqrt{2}h(z)$. In addition, $g'(z) > 0$, and therefore the function is strictly monotonically increasing, so its inverse g^{-1} is well defined. Finally, the relation between the derivatives also yields the property $g(z_2) - g(z_1) \leq \sqrt{2}(h(z_2) - h(z_1))$ for any $z_1 \leq z_2$ in $[0, 1]$.

Next, we bound $f(A; x + \delta) - f(A; x)$, for any δ such that $\delta_i \leq \min\left\{u_i\sqrt{x_i(1 - x_i)}, 1 - x_i\right\}$. The bound can be achieved using the gradient theorem:

$$f(A; x + \delta) - f(A; x) = \int_x^{x+\delta} \nabla f(A; y) \cdot dy = \int_0^1 \sum_{i \in A} \frac{\partial f(A; r(t))}{\partial x_i} r_i'(t)dt \ , \tag{22}$$

for a parameterization $r(t)$ such that $r_i(0) = x_i$ and $r_i(1) = x_i + \delta_i$. Specifically, we choose the parameterization to be

$$r_i(t) = g^{-1}([g(x_i + \delta_i) - g(x_i)]t + g(x_i)) \ ,$$

and thus its gradient is

$$r_i'(t) = \frac{g(x_i + \delta_i) - g(x_i)}{g'(r_i(t))} = (g(x_i + \delta_i) - g(x_i))\sqrt{r_i(t)(1 - r_i(t))} \ .$$

16

Substituting back into (22) yields

$$f(A; x + \delta) - f(A; x) = \int_0^1 \sum_{i \in A} (g(x_i + \delta_i) - g(x_i)) \frac{\partial f(A; r(t))}{\partial x_i} \sqrt{r_i(t)(1 - r_i(t))} dt$$

$$\leq \int_0^1 \sqrt{\sum_{i \in A} (g(x_i + \delta_i) - g(x_i))^2} \sqrt{\sum_{i \in A} \left(\frac{\partial f(A; r(t))}{\partial x_i} \right)^2 r_i(t)(1 - r_i(t))} dt \quad (23)$$

$$\leq \gamma_g \sqrt{\sum_{i \in A} (g(x_i + \delta_i) - g(x_i))^2} \ .$$

The first inequality is due to Cauchy Schwarz and the second uses the definition of the Gini-weighted smoothness (1). All that's left is bounding the differences between different values of g. Assume w.l.o.g. that $x_i \neq 0, 1$, since otherwise $\delta_i = 0$ and the difference is 0. To calculate the bound, we exploit the relation between g and h, and calculate differences over h in three different cases:

- If $x_i, x_i + \delta_i \leq \frac{1}{2}$, then

$$h(x_i + \delta_i) - h(x_i) = 2\sqrt{x_i + \delta_i} - 2\sqrt{x_i} = 2\sqrt{x_i} \left(\sqrt{1 + \frac{\delta_i}{x_i}} - 1 \right) \leq 2\sqrt{x_i} \frac{\delta_i}{2x_i}$$

$$= \frac{\delta_i}{\sqrt{x_i}} \leq u_i \ ,$$

  where the first inequality is since $\sqrt{1 + a} \leq 1 + \frac{a}{2}$ and the second is due to $\delta_i \leq u_i \sqrt{x_i}$.

- If $x_i, x_i + \delta_i \geq \frac{1}{2}$, then

$$h(x_i + \delta_i) - h(x_i) = 2\sqrt{1 - x_i} - 2\sqrt{1 - x_i - \delta_i} = 2\sqrt{1 - x_i} \left(1 - \sqrt{1 - \frac{\delta_i}{1 - x_i}} \right)$$

$$\leq 2\sqrt{1 - x_i} \frac{\delta_i}{1 - x_i} = 2 \frac{\delta_i}{\sqrt{1 - x_i}} \leq 2u_i \ ,$$

  where the first inequality is due to $1 - \sqrt{1 - a} \leq a$ for $a \in [0, 1]$ and the second is since $\delta_i \leq u_i \sqrt{1 - x_i}$.

- If $x_i \leq \frac{1}{2}$ and $x_i + \delta_i \geq \frac{1}{2}$, then

$$h(x_i + \delta_i) - h(x_i) = \left(h(x_i + \delta_i) - h(\frac{1}{2}) \right) + \left(h(\frac{1}{2}) - h(x_i) \right)$$

$$\overset{(1)}{\leq} 2 \frac{x_i + \delta_i - 1/2}{\sqrt{1 - 1/2}} + \frac{1/2 - x_i}{\sqrt{x_i}}$$

$$\overset{(2)}{\leq} 2 \frac{\delta_i}{\sqrt{x_i}} + \frac{\delta_i}{\sqrt{x_i}} \leq 3u_i \ .$$

  (1) uses the bound calculated for the first two cases and (2) takes advantage of the relation $x_i \leq \frac{1}{2}$ for the first term and $x_i + \delta_i \geq \frac{1}{2}$ for the second one.

17

To summarize, for any $x_i, x_i + \delta_i \in [0, 1]$, we have

$$g(x_i + \delta_i) - g(x_i) \leq \sqrt{2}[h(x_i + \delta_i) - h(x_i)] \leq 3\sqrt{2}u_i \ .$$

Substituting back into (23) leads to the bound

$$f(A; x + \delta) - f(A; x) \leq \gamma_g \sqrt{\sum_{i \in A}(g(x_i + \delta_i) - g(x_i))^2} \leq 3\sqrt{2}\gamma_g \sqrt{\sum_{i \in A} u_i^2} \ . \tag{24}$$

Next, we aim to bound $f(A; x + \epsilon) - f(A; x + \delta)$. Fortunately, this term can be easily bounded using the gradient theorem as follows:

$$
\begin{aligned}
f(A; x + \epsilon) - f(A; x + \delta) &= \int_{x+\delta}^{x+\epsilon} \nabla f(A; y) \cdot dy \\
&\leq \sup_y \{\|\nabla f(A; y)\|_\infty\} \sum_{i \in A}(\epsilon_i - \delta_i) \leq \gamma_\infty \sum_{i \in A} v_i
\end{aligned}
\tag{25}
$$

and combining (24) and (25) concludes the proof.

∎

Appendix B. Proof of Lemma 7

Lemma 7 *Let $r_i(A; p)$ be Gini-smooth functions with parameters γ_g, γ_∞, and define*

$$c_t(A_t) = c_1 \sqrt{\sum_{j \in A_t} \frac{\log t}{N_j(t-1)}} + c_2 \sum_{j \in A_t} \frac{\log t}{N_j(t-1)} \tag{12}$$

for $c_1 = 12\sqrt{6}\gamma_g$ and $c_2 = 34\gamma_\infty$. The regret of Algorithm 1, when used with (α, β) approximation oracle, can be bounded by

$$R(T) \leq \mathbb{E}\left[\sum_{t=L+1}^T \Delta_{A_t} \mathbb{1}\{\Delta_{A_t} \leq \bar{M}c_t(A_t)\}\right] + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right). \tag{13}$$

Proof

By the definition of the expected regret,

$$
\begin{aligned}
R(T) &= \alpha\beta r_{\max}T - \sum_{t=1}^T \mathbb{E}[r(A_t; p)] = \sum_{t=1}^T \mathbb{E}[\Delta_{A_t}] - \alpha(1-\beta)r_{\max}T \\
&\leq L\max_A \Delta_A + \mathbb{E}\left[\sum_{t=L+1}^T \Delta_{A_t} \mathbb{1}\{\mathbb{H}_t\}\right] + \\
&\quad \mathbb{E}\left[\sum_{t=L+1}^T \Delta_{A_t} \mathbb{1}\{\bar{\mathbb{H}}_t\}\right] - \alpha(1-\beta)r_{\max}T,
\end{aligned}
\tag{26}
$$

where $\mathbb{H}_t = \mathbb{H}_t^p \cup \mathbb{H}_t^V$ is the event in which the concentration bounds fail to hold at time t and $\mathbb{H}_t^p, \mathbb{H}_t^V$ are defined in Equations (10) and (11). $\bar{\mathbb{H}}_t$ is the complementary event to $\mathbb{H}_t$. The first term in (26) is the regret due to the initial sampling, that forces sampling each base arm at least once, on which we apply the bound $\Delta_A \leq \Delta_{\max}$. For the rest of the terms, the initial sampling stage allows us to assume that all of the arms were sampled at least once. The second term can be bounded using Empirical Bernstein (Lemma 1) and Bernstein inequality (7) as follows:

$$\mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\mathbb{H}_t\}\right]$$

$$\leq \mathbb{E}\left[\Delta_{\max} \sum_{t=L+1}^{T} \mathbb{1}\{\mathbb{H}_t\}\right]$$

$$\overset{(1)}{\leq} \Delta_{\max} \sum_{t=L+1}^{T} \left(\Pr\{\mathbb{H}_t^p\} + \Pr\{\mathbb{H}_t^V\}\right)$$

$$= \Delta_{\max} \sum_{t=L+1}^{T} \Pr\left\{\exists i,j : |\hat{p}_{ij}(t-1) - p_{ij}| > \sqrt{\frac{6\log t \hat{V}_{ij}(t-1)}{N_j(t-1)}} + \frac{9\log t}{N_j(t-1)}\right\} +$$

$$\Delta_{\max} \sum_{t=L+1}^{T} \Pr\left\{\exists i,j : \hat{V}_{ij}(t-1) > 2p_{ij}(1-p_{ij}) + 3.5\frac{\log t}{N_j(t-1)}\right\}$$

$$\overset{(2)}{\leq} \Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L} \Pr\left\{|\hat{p}_{ij}(t-1) - p_{ij}| > \sqrt{\frac{6\log t \hat{V}_{ij}(t-1)}{N_j(t-1)}} + \frac{9\log t}{N_j(t-1)}\right\} +$$

$$\Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L} \Pr\left\{\hat{V}_{ij}(t-1) > 2p_{ij}(1-p_{ij}) + 3.5\frac{\log t}{N_j(t-1)}\right\}$$

$$\overset{(3)}{\leq} \Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L}\sum_{s=1}^{t} \Pr\left\{|\hat{p}_{ij}(t-1) - p_{ij}| > \sqrt{\frac{6\log t \hat{V}_{ij}(t-1)}{s}} + \frac{9\log t}{s}, N_j(t-1) = s\right\} +$$

$$\Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L}\sum_{s=1}^{t} \Pr\left\{\hat{V}_{ij}(t-1) > 2p_{ij}(1-p_{ij}) + 3.5\frac{\log t}{s}, N_j(t-1) = s\right\}$$

$$\overset{(4)}{\leq} \Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L}\sum_{s=1}^{t} 3t^{-3} + \Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L}\sum_{s=1}^{t} t^{-3}$$

$$= \Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L} 3t^{-2} + \Delta_{\max} \sum_{t=L+1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{L} t^{-2}$$

$$\leq 4\Delta_{\max}ML\frac{\pi^2}{6} \tag{27}$$

Inequalities $(1) - (3)$ are simple union bounds: (1) is a union bound on $\mathbb{H}_t^p$ and $\mathbb{H}_t^V$ and on different times. (2) is on different functions $i \in [M]$ and arms $j \in [L]$. (3) is a union bound on different values of $N_j(t-1)$, ranging in $s \in [t]$. (4) uses Empirical Bernstein (Lemma 1) with $x = 3\log t$ for $\hat{p}_{ij}(t-1)$ and Bernstein's inequality (7) for $\hat{V}_{ij}(t-1)$ with $\delta = t^{-3}$.

Next, we bound the difference Δ_{A_t} under the event $\bar{\bar{\mathbb{H}}}_t$. Note that under $\bar{\bar{\mathbb{H}}}_t^p$, and for any i, j, it holds that $p_{ij} \leq q_{ij}(t)$ and

$$q_{ij}(t) \leq p_{ij} + 2\left(\sqrt{\frac{6 \log t \hat{V}_{ij}(t-1)}{N_j(t-1)}} + \frac{9 \log t}{N_j(t-1)} \right) \ .$$

In addition, since $\bar{\bar{\mathbb{H}}}_t^V$ also occurs, we can further bound q_{ij} by

$$q_{ij}(t) \leq p_{ij} + 2\left(\sqrt{\frac{6 \log t \left(2 p_{ij}(1 - p_{ij}) + 3.5 \frac{\log t}{N_j(t-1)}\right)}{N_j(t-1)}} + \frac{12 \log t}{N_j(t-1)} \right)$$

$$\leq p_{ij} + 2\left(\sqrt{\frac{12 \log t}{N_j(t-1)} p_{ij}(1 - p_{ij})} + \sqrt{21} \frac{\log t}{N_j(t-1)} + \frac{12 \log t}{N_j(t-1)} \right)$$

$$\leq p_{ij} + 4\sqrt{3} \sqrt{\frac{\log t}{N_j(t-1)} p_{ij}(1 - p_{ij})} + \frac{34 \log t}{N_j(t-1)} \ , \tag{28}$$

where we used the inequality $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$.

Notice that $q_{ij}(t)$ is of the form $q_{ij}(t) - p_{ij} \leq u_j \sqrt{p_{ij}(1 - p_{ij})} + v_j$, for $u_j = 4\sqrt{3}\sqrt{\frac{\log t}{N_j(t-1)}}$ and $v_j = 34 \frac{\log t}{N_j(t-1)}$. We can thus apply Lemma 6 to bound the difference between the optimistic reward with the UCB index q and the real reward parameters p. Combined with the properties of the approximation oracle, it is possible to estimate the error between the optimal action and the selected action, under the assumption that the oracle succeeded in his approximation (an event which we denote by $\mathcal{F}_t$)

$$\alpha r(A^*; p) \overset{(1)}{\leq} \alpha r(A^*; q) \overset{(2)}{\leq} r(A_t; q) = \sum_{i=1}^{M} w_i r_i(A_t; q)$$

$$\overset{(3)}{\leq} \sum_{i=1}^{M} w_i \left(r_i(A_t; p) + 12\sqrt{6}\gamma_g \sqrt{\sum_{j \in A_t} \frac{\log t}{N_j(t-1)}} + 34\gamma_\infty \sum_{j \in A_t} \frac{\log t}{N_j(t-1)} \right)$$

$$= r(A_t; p) + \bar{M}\left(12\sqrt{6}\gamma_g \sqrt{\sum_{j \in A_t} \frac{\log t}{N_j(t-1)}} + 34\gamma_\infty \sum_{j \in A_t} \frac{\log t}{N_j(t-1)} \right)$$

$$= r(A_t; p) + \bar{M} c_t(A_t)$$

where (1) is due to the monotonicity of r and $p \leq q$, (2) is from the properties of the approximation oracle and (3) applies Lemma 6. Consequentially, the sub-optimality gap of action A_t under the events $\bar{\bar{\mathbb{H}}}_t$ and $\mathcal{F}_t$ can be bounded by

$$\Delta_{A_t} = \alpha r_{\max} - r(A_t; p) = \alpha r(A^*; p) - r(A_t; p) \leq \bar{M} c_t(A_t) \ ,$$

which in turn, allows us to bound the third term of (26) by

$$\mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\bar{\mathbb{H}}_t\}\right] \leq \mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\bar{\mathbb{H}}_t, \mathcal{F}_t\}\right] + \mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\bar{\mathcal{F}}_t\}\right]$$

$$\leq \mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\Delta_{A_t} \leq \bar{M} c_t(A_t)\}\right] + (1 - \beta)\alpha r_{\max} T \ . \quad (29)$$

Substituting equations (27) and (29) into (26) leads to the desired result:

$$R(T) \leq L\Delta_{\max} + 2\Delta_{\max} M L \frac{\pi^2}{6} + \mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\Delta_{A_t} \leq \bar{M} c_t(A_t)\}\right]$$

■

Appendix C. Proof of Lemma 9

Lemma 9 *If $g(K, \Delta_{A_t}) = \left(864\gamma_g^2 + 68\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}\right)\bar{M}^2 K\ell$, and if $\Delta_{A_t} \leq \bar{M} c_t(A_t)$, then $\mathbb{G}_t$ occurs.*

Proof

Assume in contradiction that both $\Delta_{A_t} \leq \bar{M} c_t(A_t)$ and $\bar{\mathbb{G}}_t$ occur. From the definition of $c_t(A_t)$ and Lemma 8, we get

$$\Delta_{A_t} \leq \bar{M} c_t(A_t) = c_1 \bar{M} \sqrt{\sum_{j \in A_t} \frac{\log t}{N_j(t-1)}} + \bar{M} c_2 \sum_{j \in A_t} \frac{\log t}{N_j(t-1)}$$

$$= 12\sqrt{6}\gamma_g \bar{M} \sqrt{\sum_{j \in A_t} \frac{\log t}{N_j(t-1)}} + 34\gamma_\infty \bar{M} \sum_{j \in A_t} \frac{\log t}{N_j(t-1)}$$

$$< 12\sqrt{6}\gamma_g \bar{M} \sqrt{\frac{K\ell\Delta_{A_t}^2}{g(K, \Delta_{A_t})}} + 34\gamma_\infty \bar{M} \frac{K\ell\Delta_{A_t}^2}{g(K, \Delta_{A_t})}$$

$$= \frac{12\sqrt{6}\gamma_g}{\sqrt{864\gamma_g^2 + 68\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}}} \Delta_{A_t} + \frac{34\gamma_\infty}{\bar{M}\left(864\gamma_g^2 + 68\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}\right)} \Delta_{A_t}^2$$

$$= \frac{12\sqrt{6}\gamma_g \sqrt{864\gamma_g^2 + 68\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}} + 34\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}}{\left(864\gamma_g^2 + 68\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}\right)} \Delta_{A_t}$$

$$\overset{(*)}{\leq} \frac{864\gamma_g^2 + 34\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}} + 34\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}}{\left(864\gamma_g^2 + 68\frac{\gamma_\infty \Delta_{A_t}}{\bar{M}}\right)} \Delta_{A_t}$$

$$= \Delta_{A_t} \ ,$$

where $(*)$ is due to $\sqrt{a(a+b)} \leq a + \frac{b}{2}$. We got $\Delta_{A_t} < \Delta_{A_t}$, which is a contradiction, and therefore, if $\Delta_{A_t} \leq \bar{M} c_t(A_t)$, then $\bar{\mathbb{G}}_t$ cannot occur. ■

Appendix D. Proof of Lemma 10

Lemma 10 *The regret from the event $\Delta_{A_t} \leq \bar{M}c_t(A_t)$ can be bounded by*

$$\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\Delta_{A_t} \leq \bar{M}c_t(A_t)\}$$

$$\leq \left[1728\gamma_g^2 \bar{M}^2 \sum_{j=1}^{L} \frac{1}{\Delta_{j,\min}} + 68\gamma_\infty \bar{M} \sum_{j=1}^{L}\left(1 + \log\frac{\Delta_{j,\max}}{\Delta_{j,\min}}\right)\right]\left(\sum_{k=1}^{k_0} \frac{a_k}{b_k}\right)\ell \log T \ , \quad (15)$$

and if $a_k = b_k = 0.2^k$,

$$\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\{\Delta_{A_t} \leq \bar{M}c_t(A_t)\}$$

$$\leq \left[8640\gamma_g^2 \bar{M}^2 \sum_{j=1}^{L} \frac{1}{\Delta_{j,\min}} + 340\gamma_\infty \bar{M} \sum_{j=1}^{L}\left(1 + \log\frac{\Delta_{j,\max}}{\Delta_{j,\min}}\right)\right]\left\lceil\frac{\log K}{1.61}\right\rceil^2 \log T \ . \quad (16)$$

Proof

Recall that $\mathbb{G}_t^k = \{\{|S_t^k| \geq Kb_k\} \cap \{\forall n < k, |S_t^n| < Kb_n\}\}$. Specifically, if $\mathbb{G}_t^k$ occurs, then at least Kb_k arms were sampled less than $a_k \frac{g(K,\Delta_{A_t})\log t}{\Delta_{A_t}^2}$ times. Denote the sub-event of $\mathbb{G}_t^k$ in which j is one of these arms by $\mathbb{G}_t^{k,j} = \mathbb{G}_t^k \cap \left\{j \in A_t, N_j(t-1) \leq a_k \frac{g(K,\Delta_{A_t})\log t}{\Delta_{A_t}^2}\right\}$. Therefore, if $\mathbb{G}_t^k$ occurs, there are at least Kb_k arms for which $\mathbb{G}_t^{k,j}$ occur, and

$$\mathbb{1}\left\{\mathbb{G}_t^k\right\} \leq \frac{1}{Kb_k}\sum_{j=1}^{L}\mathbb{1}\left\{\mathbb{G}_t^{k,j}\right\} \ .$$

Next, we apply Lemma 9, and bound the regret from the event $\Delta_{A_t} \leq \bar{M}c_t(A_t)$ by

$$\sum_{t=L+1}^{T} \Delta_{A_t}\mathbb{1}\{\Delta_{A_t} \leq \bar{M}c_t(A_t)\} \leq \sum_{t=L+1}^{T}\sum_{k=1}^{k_0}\Delta_{A_t}\mathbb{1}\left\{\mathbb{G}_t^k\right\} \leq \sum_{t=L+1}^{T}\sum_{k=1}^{k_0}\sum_{j=1}^{L}\frac{\Delta_{A_t}}{Kb_k}\mathbb{1}\left\{\mathbb{G}_t^{k,j}\right\} \ .$$

Denote the number of possible positive gaps of batches that contain arm j by D_j, and assume $\Delta_{j,1} >, \cdots > \Delta_{j,D_j} = \Delta_{j,\min} > 0$ with $\Delta_{j,0} = \infty$. We decompose the regret as

$$\sum_{t=L+1}^{T} \Delta_{A_t}\mathbb{1}\{\Delta_{A_t} \leq \bar{M}c_t(A_t)\} \leq \sum_{t=L+1}^{T}\sum_{k=1}^{k_0}\sum_{j=1}^{L}\sum_{n=1}^{D_j}\frac{\Delta_{j,n}}{Kb_k}\mathbb{1}\left\{\mathbb{G}_t^{k,j}, \Delta_{A_t} = \Delta_{j,n}\right\}$$

$$\leq \sum_{t=L+1}^{T}\sum_{k=1}^{k_0}\sum_{j=1}^{L}\sum_{n=1}^{D_j}\frac{\Delta_{j,n}}{Kb_k}\mathbb{1}\left\{j \in A_t, N_j(t-1) \leq a_k\frac{g(K,\Delta_{A_t})\log t}{\Delta_{j,n}^2}, \Delta_{A_t} = \Delta_{j,n}\right\} \ . \quad (30)$$

22

Denote $\theta_k = 864\gamma_g^2 \bar{M}^2 K \ell a_k \log T$ and $\mu_k = 68\gamma_\infty \bar{M} K \ell a_k \log T$. Under these notations

$$\sum_{t=L+1}^{T} \sum_{n=1}^{D_j} \Delta_{j,n} \mathbb{1}\left\{ j \in A_t, N_j(t-1) \le a_k \frac{g(K, \Delta_{A_t})\log t}{\Delta_{j,n}^2}, \Delta_{A_t} = \Delta_{j,n} \right\}$$

$$\overset{(1)}{\le} \sum_{t=L+1}^{T} \sum_{n=1}^{D_j} \Delta_{j,n} \mathbb{1}\left\{ j \in A_t, N_j(t-1) \le \frac{\theta_k}{\Delta_{j,n}^2} + \frac{\mu_k}{\Delta_{j,n}}, \Delta_{A_t} = \Delta_{j,n} \right\}$$

$$\overset{(2)}{\le} \sum_{t=L+1}^{T} \sum_{n=1}^{D_j} \sum_{p=1}^{n} \Delta_{j,n} \mathbb{1}\left\{ j \in A_t, N_j(t-1) \in \left(\frac{\theta_k}{\Delta_{j,p-1}^2} + \frac{\mu_k}{\Delta_{j,p-1}}, \frac{\theta_k}{\Delta_{j,p}^2} + \frac{\mu_k}{\Delta_{j,p}} \right], \Delta_{A_t} = \Delta_{j,n} \right\}$$

$$\overset{(3)}{\le} \sum_{t=L+1}^{T} \sum_{n=1}^{D_j} \sum_{p=1}^{n} \Delta_{j,p} \mathbb{1}\left\{ j \in A_t, N_j(t-1) \in \left(\frac{\theta_k}{\Delta_{j,p-1}^2} + \frac{\mu_k}{\Delta_{j,p-1}}, \frac{\theta_k}{\Delta_{j,p}^2} + \frac{\mu_k}{\Delta_{j,p}} \right], \Delta_{A_t} = \Delta_{j,n} \right\}$$

$$\overset{(4)}{\le} \sum_{t=L+1}^{T} \sum_{n=1}^{D_j} \sum_{p=1}^{D_j} \Delta_{j,p} \mathbb{1}\left\{ j \in A_t, N_j(t-1) \in \left(\frac{\theta_k}{\Delta_{j,p-1}^2} + \frac{\mu_k}{\Delta_{j,p-1}}, \frac{\theta_k}{\Delta_{j,p}^2} + \frac{\mu_k}{\Delta_{j,p}} \right], \Delta_{A_t} = \Delta_{j,n} \right\}$$

$$\overset{(5)}{\le} \sum_{t=L+1}^{T} \sum_{p=1}^{D_j} \Delta_{j,p} \mathbb{1}\left\{ j \in A_t, N_j(t-1) \in \left(\frac{\theta_k}{\Delta_{j,p-1}^2} + \frac{\mu_k}{\Delta_{j,p-1}}, \frac{\theta_k}{\Delta_{j,p}^2} + \frac{\mu_k}{\Delta_{j,p}} \right], \Delta_{A_t} > 0 \right\}$$

$$\overset{(6)}{\le} \frac{\theta_k}{\Delta_{j,1}} + \mu_k + \sum_{p=2}^{D_j} \Delta_{j,p} \theta_k \left(\frac{1}{\Delta_{j,p}^2} - \frac{1}{\Delta_{j,p-1}^2} \right) + \sum_{p=2}^{D_j} \Delta_{j,p} \mu_k \left(\frac{1}{\Delta_{j,p}} - \frac{1}{\Delta_{j,p-1}} \right)$$

$$= \frac{\theta_k}{\Delta_{j,D_j}} + +\mu_k + \theta_k \sum_{p=1}^{D_j-1} \frac{\Delta_{j,p} - \Delta_{j,p+1}}{\Delta_{j,p}^2} + \mu_k \sum_{p=1}^{D_j-1} \frac{\Delta_{j,p} - \Delta_{j,p+1}}{\Delta_{j,p}}$$

$$\le \frac{\theta_k}{\Delta_{j,D_j}} + \mu_k + \theta_k \int_{\Delta_{j,D_j}}^{\Delta_{j,1}} x^{-2} dx + \mu_k \int_{\Delta_{j,D_j}}^{\Delta_{j,1}} x^{-1} dx$$

$$\le \frac{2\theta_k}{\Delta_{j,D_j}} + \mu_k + \mu_k \log \frac{\Delta_{j,1}}{\Delta_{j,D_j}}$$

$$= \frac{2\theta_k}{\Delta_{j,\min}} + \mu_k \left(1 + \log \frac{\Delta_{j,\max}}{\Delta_{j,\min}} \right).$$

We added numbering through the first lines, as changes between consecutive lines can be hard to discern. In (1), we bounded $a_k g(K, \Delta_{A_t})\log t \le \theta_k + \mu_k \Delta_{A_t}$. In (2), we divided the interval $N_j(t-1) \le \frac{\theta_k}{\Delta_{j,n}^2}$ into non overlapping sub-interval and used union bound. Next, we replaced $\Delta_{j,n}$ by larger $\Delta_{j,p}$ in (3), and extended the internal sum to D_j in (4). Finally, in (5) we replaced the sum over specific positive gaps by the event $\Delta_{A_t} > 0$ and in (6) we bounded the maximal number of times that each of the indicators can be nonzero by he length of the interval. The rest of the lines involve reordering and bounding a summation by an integral. Substituting back into (30) gives the first desired result:

$$\sum_{t=L+1}^{T} \Delta_{A_t} \mathbb{1}\left\{ \Delta_{A_t} \le \bar{M} c_t(A_t) \right\} \le \sum_{k=1}^{k_0} \sum_{j=1}^{L} \frac{1}{K b_k} \left(\frac{2\theta_k}{\Delta_{j,\min}} + \mu_k \left(1 + \log \frac{\Delta_{j,\max}}{\Delta_{j,\min}} \right) \right)$$

$$\le \left[1728\gamma_g^2 \bar{M}^2 \sum_{j=1}^{L} \frac{1}{\Delta_{j,\min}} + 68\gamma_\infty \bar{M} \sum_{j=1}^{L} \left(1 + \log \frac{\Delta_{j,\max}}{\Delta_{j,\min}} \right) \right] \left(\sum_{k=1}^{k_0} \frac{a_k}{b_k} \right) \ell \log T .$$

Similarly to appendix C of (Degenne and Perchet, 2016), choosing $a_k = b_k = b^k$ for $b = 0.2$

$$\left(\sum_{k=1}^{k_0} \frac{a_k}{b_k}\right)\ell = \left(\sum_{k=1}^{k_0} \frac{a_k}{b_k}\right)\left[\sum_{k=1}^{k_0} \frac{b_{k-1}-b_k}{a_k} + \frac{b_0}{a_0}\right] \le \frac{1}{b}\left\lceil \frac{\log K}{\log 1/b}\right\rceil^2 \le 5\left\lceil \frac{\log K}{1.61}\right\rceil^2,$$

which concludes the second part of the results. ∎

Appendix E. Proof of Corollary 4

Corollary 4 *Let $r_i(A; p_i)$ be monotonic Gini-smooth reward functions with smoothnsess parameters γ_∞ and γ_g, and let $r(A; p) = \sum_{i=1}^{M} w_i r_i(A; p_i)$ be the reward function. For any $T \ge 1$, the expected approximation regret of BC-UCB with (α,β)-approximation oracle can be bounded by*

$$R(T) \le 2\sqrt{8640}\gamma_g \bar{M}\left\lceil \frac{\log K}{1.61}\right\rceil \sqrt{LT\log T} + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right)$$

$$+ 340\gamma_\infty \bar{M}L\left\lceil \frac{\log K}{1.61}\right\rceil^2 \log T\left(2 + \log \frac{\Delta_{\max}T}{340\gamma_\infty \bar{M}L\left\lceil \frac{\log K}{1.61}\right\rceil^2 \log T}\right). \tag{5}$$

Proof

We start from Lemma 7, but divide the first term of the regret into large gaps $\Delta_{A_t} \ge \Delta$ and small gaps $\Delta_{A_t} \le \Delta$. We then use Lemma 10 only for the large gaps, and bound the regret of the smaller gaps by the trivial bound ΔT:

$$R(T) \le \mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t}\mathbb{1}\{\Delta_{A_t} \le \bar{M}c_t(A_t)\}\right] + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right)$$

$$\le \mathbb{E}\left[\sum_{t=L+1}^{T} \Delta_{A_t}\mathbb{1}\{\Delta_{A_t} \le \bar{M}c_t(A_t), \Delta_{A_t} \ge \Delta\}\right] + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right)$$

$$+ \mathbb{E}\left[\sum_{t=1}^{T} \Delta_{A_t}\mathbb{1}\{\Delta_{A_t} < \Delta\}\right]$$

$$\le \left[8640\gamma_g^2\bar{M}^2 \sum_{j:\Delta_{j,\min}>\Delta} \frac{1}{\Delta_{j,\min}} + 340\gamma_\infty\bar{M} \sum_{j:\Delta_{j,\min}>\Delta}\left(1 + \log \frac{\Delta_{j,\max}}{\Delta_{j,\min}}\right)\right]\left\lceil \frac{\log K}{1.61}\right\rceil^2 \log T$$

$$+ L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right) + T\Delta$$

$$\le \left[8640\gamma_g^2\bar{M}^2 \frac{L}{\Delta} + 340\gamma_\infty\bar{M}L\left(1 + \log \frac{\Delta_{\max}}{\Delta}\right)\right]\left\lceil \frac{\log K}{1.61}\right\rceil^2 \log T$$

$$+ L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3}\right) + T\Delta.$$

Next, set $\Delta = \sqrt{\frac{u_1 \log T}{T}} + \frac{u_2 \log T}{T}$, for $u_1 = 8640\gamma_g^2\bar{M}^2 L\left\lceil \frac{\log K}{1.61}\right\rceil^2$ and $u_2 = 340\gamma_\infty\bar{M}L\left\lceil \frac{\log K}{1.61}\right\rceil^2$. Substituting this value gives the desired results:

$$
R(T) \leq \left[\frac{8640\gamma_g^2 \bar{M}^2 L}{\sqrt{\frac{u_1 \log T}{T} + \frac{u_2 \log T}{T}}} + 340\gamma_\infty \bar{M} L \log \frac{\Delta_{\max}}{\sqrt{\frac{u_1 \log T}{T} + \frac{u_2 \log T}{T}}} \right] \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T
$$

$$
+ 340\gamma_\infty \bar{M} L \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3} \right) + T\left(\sqrt{\frac{u_1 \log T}{T}} + \frac{u_2 \log T}{T} \right)
$$

$$
\leq \left[\frac{8640\gamma_g^2 \bar{M}^2 L}{\sqrt{\frac{u_1 \log T}{T}}} + 340\gamma_\infty \bar{M} L \log \frac{\Delta_{\max}}{\frac{u_2 \log T}{T}} \right] \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T
$$

$$
+ 340\gamma_\infty \bar{M} L \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3} \right) + \sqrt{u_1 T \log T} + u_2 \log T
$$

$$
= 2\sqrt{8640}\gamma_g \bar{M} \left\lceil \frac{\log K}{1.61} \right\rceil \sqrt{LT \log T} + L\Delta_{\max}\left(1 + M\frac{2\pi^2}{3} \right)
$$

$$
+ 340\gamma_\infty \bar{M} L \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T \left(2 + \log \frac{\Delta_{\max}T}{340\gamma_\infty \bar{M} L \left\lceil \frac{\log K}{1.61} \right\rceil^2 \log T} \right)
$$

$\blacksquare$