CONTEXTUAL COMBINATORIAL VOLATILE MULTI-ARMED BANDITS IN COMPACT CONTEXT SPACES

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

By

Andi Nika

July 2021

CONTEXTUAL COMBINATORIAL VOLATILE MULTI-ARMED BANDITS IN COMPACT CONTEXT SPACES

By Andi Nika

July 2021

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Cem Tekin(Advisor)

Çağın Ararat

Emre Özkan

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

CONTEXTUAL COMBINATORIAL VOLATILE MULTI-ARMED BANDITS IN COMPACT CONTEXT SPACES

Andi Nika
M.S. in Electrical and Electronics Engineering
Advisor: Cem Tekin
July 2021

We consider the contextual combinatorial volatile multi-armed bandit (CCV-MAB) problem in compact context spaces, simultaneously taking into consideration all of its individual features, thus providing a general framework for solving a wide range of practical problems. We solve CCV-MAB using two approaches. First, we use the so called adaptive discretization technique which sequentially partitions the context space $\mathcal{X}$ into 'regions of similarity' and stores similar statistics corresponding to such regions. Under monotonicity of the expected reward and mild continuity assumptions, for both the expected reward and the expected base arm outcomes, we propose *Adaptive Contextual Combinatorial Upper Confidence Bound* (ACC-UCB), an online learning algorithm that uses adaptive discretization and incurs $\tilde{O}(T^{(\bar{D}+1)/(\bar{D}+2)+\epsilon})$ regret for any $\epsilon > 0$, where $\bar{D}$ represents the approximate optimality dimension related to $\mathcal{X}$. This dimension captures both the benignness of the base arm arrivals and the structure of the expected reward. Second, we impose a Gaussian process (GP) structure on the expected base arms outcomes and thus, using the smoothness of the GP posterior, eliminate the need for adaptive discretization. We propose *Optimistic Combinatorial Learning and Optimization with Kernel Upper Confidence Bounds* (O'CLOK-UCB) which incurs $\tilde{O}(K\sqrt{T\bar{\gamma}_T})$ regret, where $\bar{\gamma}_T$ is the maximum information gain associated with the set of base arm contexts that appeared in the first T rounds and K here is the maximum cardinality of any feasible super arm over all rounds. For both methods, we provide experimental results which conclude in the superiority of ACC-UCB over the previous state-of-the-art and of O'CLOCK-UCB over ACC-UCB.

Keywords: multi-armed bandit, contextual combinatorial bandit, volatile bandit, adaptive discretization, Gaussian processes.

ÖZET

TIKIZ BAĞLAM UZAYLARINDA BAĞLAMSAL BİRLEŞİMSEL DEĞİŞKEN ÇOK-KOLLU HAYDUT

Andi Nika
Elektrik ve Elektronik Mühendislik, Yüksek Lisans
Tez Danışmanı: Cem Tekin
July 2021

Tıkız bağlam uzaylarında bağlamsal birleşimsel çok-kollu değişken haydut problemi (CCV-MAB), eşzamanlı olarak tüm özgün özellikleri göz önünde bulundurularak, geniş bir yelpazede pratik problemleri çözmek için genel bir çerçeve sunarak ele alınmıştır. CCV-MAB iki yaklaşım kullanılarak çözülmüştür. İlk olarak, bağlam uzayı $\mathcal{X}$'i 'benzer bölgelere' sıralı ayrıştıran ve bu bölgelere karşılık gelen benzer istatistikleri kaydeden uyarlanır ayrıklaştırma adlı yöntem kullanılmıştır. Beklenen ödül monotonluğu ve zayıf süreklilik varsayımları altında, beklenen ödül ve beklenen taban kolu sonuçları için, uyarlanır ayrıklaştırma kullanan ve tüm $\epsilon > 0$ için $\tilde{O}(T^{(\bar{D}+1)/(\bar{D}+2)+\epsilon})$ pişmanlığa sahip bir çevrimiçi öğrenme algoritması *Uyarlanır Bağlamsal Birleşimsel Üst Güven Sınırı* (ACC-UCB) önerilmiştir, $\bar{D}$ $\mathcal{X}$ ile ilişkili yaklaşık eniyilik boyutunu temsil etmektedir. Bu boyut taban kolu varışlarının iyicilliğini ve beklenen ödül yapısını yakalar. İkinci olarak, beklenen taban kolu sonuçları Gauss süreci (GP) kullanılarak modellenmiştir. Bu sayede GP sonsalının düzgünlüğü kullanılarak uyarlanır ayrıklaştırma ihtiyacı ortadan kaldırılmıştır. Çalışmada $\tilde{O}(K\sqrt{T\bar{\gamma}_T})$ pişmanlığa sahip *Çekirdek Üst Güven Sınırı ile İyimser Birleşimsel Öğrenme ve Eniyileme* (O'CLOK-UCB) önerilmiştir. $\bar{\gamma}_T$ ilk T turda görünen taban kolu bağlam kümesiyle ilişkili azami bilgi kazancı ve K tüm turlar arasında herhangi bir olurlu üstün kolun azami sayallığıdır. ACC-UCB'nin literatürdeki güncel algoritmalardan ve O'CLOK-UCB'un ACC-UCB'den üstünlüğü simülasyon ortamında gözlemlenmiştir.

Anahtar sözcükler: çok-kollu haydut, bağlamsal birleşimsel haydut, değişken haydut, uyarlanır ayrıklaştırma, Gauss süreci.

Acknowledgement

This thesis is a product of a three-year-long, fruitful collaboration with Assoc. Prof. Cem Tekin. His expertise, mentorship and discipline throughout this program have been highly invaluable to my professional growth. I am grateful to him for helping me understand what research is.

I want to thank Bilkent University for providing me the financial means to complete my studies. These five years in its outstanding community have been some of the most beautiful and insightful years of my life. I have substantially matured during these years and have had the chance to know some fascinating people for whose company I am deeply grateful. I am thankful to Cansın Çakır, for all those enlightening long walks; Franci Gripshi, for being a true brother in any time of need; Gerardo Tatzati, for always being on the other side of the phone; Alparslan Çelik, for all those talks in which we shared our common struggles; Umur Çetin for being a very respectful and helping flat-mate; Ela Fasllija and Asu Işbilen, for always showing their support and cheering me up.

I am indebted to my father Gani, my mother Rubjana, and my brother Geri, for their continuous encouragement throughout my education. Their love and commitment have been crucial in everything I have pursued. Having them by my side has truly been a blessing.

Finally, I want to express my gratitude to Sahra, who has been my best friend, partner, and companion throughout these last two years. Her dedication and maturity have been profoundly nurturing and encouraging throughout my journey here. I am truly grateful for all the things we came to learn together.

This work was supported by TÜBITAK project no: 215E342.

Contents

Chapter 1

Introduction

Sequential decision-making under uncertainty has been studied extensively in recent years, expanding its research domain from pure statistics to reinforcement learning. The exponential growth in computational power has made possible the application of theoretical models to real-world high-complexity problems. Whether it is web-advertisement, movie recommendations, or self-driving cars, the common denominator is the sequential nature of these problems, making the study of Reinforcement Learning (RL) necessary and valuable. Recent research has provided powerful state-of-the-art algorithms in order to approach and efficiently solve such problems. The standard performance measures for such algorithms depend on their objectives. The goal may be to converge in as few samples as possible, in which case the solution lies with algorithms that minimize sample-complexity, or to guarantee that, after some point in time, the optimal choice will always be selected, and here, algorithms which minimize regret are the suitable choice. However, whether it is regret or sample-complexity minimization, such algorithms' theoretical guarantees make their application safe and beneficial. They are, so to speak, a measure of their trustworthiness.

1.1 Preliminaries and literature review

The multi-armed bandit (MAB) is a prominent example of the sequential decision-making paradigm under uncertainty [1, 2]. In its classical version, an N-armed bandit problem is defined by the random variables $r_{s,t}$, where $1 \leq s \leq N$ is the arm index and $t \geq 1$ is the time index, for $N \in \mathbb{N}$. Basically, the learner selects arms sequentially over rounds, one at a time from a finite set, in order to maximize its cumulative reward [3–5]. Successive plays of the same arm are assumed to be independent and identically distributed across rounds. The same holds true for plays of different arms in the same round. The agent faces the challenge of exploration and exploitation as it is not aware of the arm reward distributions beforehand and observes in each round only the reward of the selected arm s_t, i.e. the realization of the random variable $r_{s_t,t}$. We denote the expected outcome associated with arm s by $\mu(s) := \mathbb{E}[r_{s,t}]$, for any $1 \leq t$. Note that the time index here does not matter, since $\mathbb{E}[r_{s,t}] = \mathbb{E}[r_{s,t'}]$, for any $1 \leq t, t'$. This follows from the fact that any two samples of the same arm in different rounds come from the same distribution, and thus have the same expected value.

1.1.1 The regret

The metric used to evaluate the performance of the learner is called the (expected) regret, which is defined as the cumulative loss of the learner in T rounds with respect to an oracle that acts optimally based on arms' reward distributions, where $T \in \mathbb{N}$ here is called the *time horizon*. In the classical N-armed bandit problem, there exists an arm s^*[1] such that $\mu(s^*) \geq \mu(s)$, for all $s \in [N]$[2]. Formally, we let $s^* := \operatorname{argmax}_{1 \leq s \leq N} \mu(s)$ be the optimal arm that the oracle selects in every round. The regret in itself is a random variable defined as

$$\mathcal{R}(T) = \sum_{t=1}^{T} r_{s^*,t} - \sum_{t=1}^{T} r_{s_t,t} \ .$$

[1] Note that s^* is not said to be unique. Without loss of generality, any such maximizer would suffice, since we are only interested in $\mu(s^*)$.

[2] $[N] := \{1, 2, \ldots, N\}$.

Note that there may exist some round $t \geq 1$ in which $r_{s^*,t} < r_{s,t}$, for some $s \in [N]$. This is perfectly normal, and actually inevitable. We are here dealing with randomness and the best we can do (in the optimal case when we know the distributions beforehand) is to go with the distribution with the highest mean, since it returns the highest rewards 'in expectation'. Thus, what we want to minimize is the expected value of $\mathcal{R}(T)$, that is,

$$R(T) := \mathbb{E}[\mathcal{R}(T)] = \mathbb{E}\left[\sum_{t=1}^{T} r_{s^*,t}\right] - \mathbb{E}\left[\sum_{t=1}^{T} r_{s_t,t}\right] = \sum_{t=1}^{T} \mu(s^*) - \sum_{t=1}^{T} \mathbb{E}\left[\mu(s_t)\right] \ ,$$

where we use the linearity of expectation to obtain the result.

There is another way in which we can decompose the expected regret in the case when T and N are both finite and the rewards are stochastic. Let us define the *suboptimality gap* of an arm s as $\Delta_s = \mu(s^*) - \mu(s)$. Furthermore, for a given round $t \leq T$, let $n_t(s) = \sum_{j=1}^{t} \mathbb{I}\left(s_j = s\right)$ denote the number of rounds arm s has been chosen up until round t, where $\mathbb{I}(\cdot)$ denotes the indicator function. Note that $\sum_{t=1}^{T} r_{s_t,t} = \sum_{t=1}^{T} \sum_{s=1}^{N} r_{s,t} \mathbb{I}\left(s_t = s\right)$, since $\sum_{s=1}^{N} \mathbb{I}(s_t = s) = 1$. Therefore, we have

$$\begin{aligned}
R(T) = \sum_{t=1}^{T} \mu(s^*) - \sum_{t=1}^{T} \mathbb{E}[r_{s_t,t}] &= \sum_{s=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu(s^*) - r_{s_t,t}\right) \mathbb{I}\left(s_t = s\right)\right] && (1.1)\\
&= \sum_{s=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\left(\left(\mu(s^*) - r_{s_t,t}\right) \mathbb{I}\left(s_t = s\right)\right)\right) \Big| s_t\right]\right] && (1.2)\\
&= \sum_{s=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\left(\mu(s^*) - \mu(s_t)\right) \mathbb{I}\left(s_t = s\right)\right] && (1.3)\\
&= \sum_{s=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\Delta_s \mathbb{I}\left(s_t = s\right)\right)] && (1.4)\\
&= \sum_{s=1}^{N} \sum_{t=1}^{T} \Delta_s \mathbb{E}\left[\mathbb{I}\left(s_t = s\right)\right] && (1.5)\\
&= \sum_{s=1}^{N} \Delta_s \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(s_t = s\right)\right] && (1.6)
\end{aligned}$$

$$= \sum_{s=1}^{N} \Delta_s \mathbb{E}\left[n_T(s)\right] \ , \tag{1.7}$$

where for (1.1) we have used the argument above and the linearity of expectation; for (1.2) we use the tower rule and condition on the random variable s_t; for (1.3) note that, conditioned on s_t, we have $\mathbb{E}[r_{s_t,t}] = \mu(s_t)$; in (1.4) we have used the definition of suboptimality gap given above; in (1.5) we take Δ_s out since it is determined, not random, and finally in (1.6) we rearrange the terms so that we get $n_T(s)$ in (1.7).

The alternative way of expressing the regret is very useful when one wants to work with suboptimality gaps and expected number of times an arm is selected. This is usually the case when we are working in a finite setting, that is, the number of arms is finite. Maximizing the cumulative reward is equivalent to minimizing the regret and that is why we tend to focus on regret minimization in bandit problems. However, when the learning agent is oblivious to the arms' distributions, he cannot select the optimal arm, at least during the first rounds. He needs to undergo a trial-and-error process in order to accurately learn their distributions. And, during this process, he needs to learn from past choices in order to choose better. But how to learn from past choices? The first thing that comes to mind is the sample average, or empirical mean, of a given arm. There is a very powerful result in probability theory that helps this part of the learning process. It is called the *Strong Law of Large Numbers*.

Theorem. *Let $X_1, X_2, \ldots, X_m$ be a sequence of independent and identically distributed random variables, for which $\mathbb{E}[X_k] = \mu_X < \infty$, for all $k \leq m$ and some $m \in \mathbb{N}$. Let $S_n = X_1 + \ldots + X_n$, for any $n \leq m$. Then we have*

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{S}_n - \mu_X| \geq \epsilon\right) = 0 \ ,$$

for any $\epsilon > 0$, where $\bar{S}_n = S_n/n$.

Intuitively, what we understand from the theorem above is that, the more (independent) samples we obtain from a certain distribution with finite mean, the closer the sample mean gets to the real one. Therefore, one approach to solve the learning problem when the distributions are unknown would be to sample as much as possible from every arm and compare the sample means. But this would be very expensive and not

efficient at all. What we seek is a method which is not just blind sampling. Surely, once we see that a certain arm is giving us really good outcomes, we may also choose to stick to it, in other words, to *exploit* it. But this in itself may cause us to stick to a suboptimal arm, since we may have not sampled enough from the optimal arm to detect it. This is the so called exploration-exploitation dilemma, and the learning agent must follow a policy[3] that addresses it.

1.1.2 The ϵ-greedy algorithm

One approach is to dedicate some rounds to exploration and some to the exploitation process. A simple bandit algorithm called ϵ-greedy [6] proceeds as follows: first, all arms are chosen once, and then, at every round, the algorithm will select the arm with the highest sample mean with probability $1 - \epsilon$ and will randomly pick an arm with probability ϵ, where $\epsilon \in (0, 1)$ is a parameter to be chosen. We will show that, if ϵ is fixed throughout all rounds, we are left with linear regret. First, note that the sample mean of a given arm s at the beginning of round t can be written as

$$\bar{\mu}_t(s) = \frac{1}{n_t(s)} \sum_{j=1}^{t} r_{s_j,j} \mathbb{I}\left(s_j = s\right) \ .$$

Now, according to the procedure that ϵ-greedy follows, we can express the random variable s_t as

$$s_t = \begin{cases} (t \bmod N) + 1 & \text{if } t \leq N \\ \operatorname{argmax}_{1 \leq s \leq N} \bar{\mu}_t(s) & \text{with probability } 1 - \epsilon, \text{ if } t > N \\ S & \text{with probability } \epsilon, \text{ if } t > N, \end{cases}$$

where S is a uniform random variable taking values in N. Note that, in the first N rounds, the ϵ-greedy policy is deterministic. Now we can use the alternative version of the regret that we stated above in order to compute it for ϵ-greedy. We need to compute

[3] A policy is defined as the rule by which an agent sequentially selects arms over rounds in a bandit game.

$\mathbb{E}[n_T(s)]$, for a given arm s and time horizon T. Note that we have

$$\begin{aligned}\mathbb{E}\left[n_T(s)\right] &= 1 + \sum_{t=N+1}^{T}\left[\epsilon \cdot \frac{1}{N} + (1-\epsilon)\,\mathbb{P}\left(\underset{1\le k\le N}{\operatorname{argmax}}\,\bar{\mu}_t(k) = s\right)\right] \\ &= 1 + \frac{\epsilon}{N}(T-N-1) + remainder \\ &\geq 1 + \frac{\epsilon}{N}(T-N-1)\,,\end{aligned}$$

since the $remainder$ term is non-negative. Now going back to the regret of ϵ-greedy, we have

$$R(T) = \sum_{s=1}^{N}\Delta_s\mathbb{E}\left[n_T(s)\right] \geq \sum_{s=1}^{N}\Delta_s\left(1 + \frac{\epsilon}{N}(T-N-1)\right) = \Omega\left(T\frac{\epsilon}{N}\sum_{s=1}^{N}\Delta_s\right)\,,$$

since we know for sure that $T > N$. This implies that the regret grows infinitely with T, since, asymptotically, the regret rate stays constant. Thus, we cannot guarantee learning here. Auer et. al. solved this problem by assigning $\epsilon_t = 1/t$ in [5] and obtained sublinear regret[4].

1.1.3 The Upper Confidence Bound method

Before introducing the Upper Confidence Bound (UCB) method, let us first recall the famous Hoeffding bounds for bounded i.i.d. random variables.

Theorem 1.1. *Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite support in $[0,1]$ and let $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ denote their empirical mean. Then we have*

$$\mathbb{P}\left(\bar{X} - \mathbb{E}[\bar{X}] \geq t\right) \leq \exp\left(-2nt^2\right)\,.$$

From the Strong Law of Large Numbers it intuitively follows that, if we are to keep a statistic for a given arm (and one that helps exploitation), it would at least consist of the sample mean. But, in order to balance exploration and exploitation, we need to add (or subtract) another term which would take care of exploration. Many algorithms

[4]When the exponent of T in the regret bounds is strictly less than 1 we call it sublinear.

have been proposed to minimize the regret by balancing exploration and exploitation. Two notable examples are upper confidence bound (UCB) based index policies [2,5,7] and Thompson sampling [1,8,9].

For each arm s, we define an index which serves as an estimate of the expected outcome associated with that arm, that is, given round t, the index $i_t(s)$ of arm s is defined as

$$i_t(s) = \bar{\mu}_t(s) + \sqrt{\frac{2\ln t}{n_t(s)}}\,.$$

The second term of the index serves as a 'confidence radius', thus giving a margin of error of the estimate that we have for the true mean of s. We will now explain the reason of using such a confidence radius.

Using Theorem 1.1 we can check how accurately this index estimates the true mean of the arm. For that we need to see whether or not we have

$$\mu(s) \in \left(\bar{\mu}_t(s) - \sqrt{\frac{2\ln t}{n_t(s)}}, \bar{\mu}_t(s) + \sqrt{\frac{2\ln t}{n_t(s)}}\right)$$

at any given round $t \geq 1$. The interval above is called the *confidence interval*. Without loss of generality, given $t \geq 1$, we take the upper bound and write

$$\begin{aligned}
1 - \mathbb{P}\left(\mu(s) < \bar{\mu}_t(s) + \sqrt{\frac{2\ln t}{n_t(s)}}\right) &= \mathbb{P}\left(\mu(s) - \bar{\mu}_t(s) \geq \sqrt{\frac{2\ln t}{n_t(s)}}\right) \\
&= \sum_{t'=1}^{t} \mathbb{P}\left(\mu(s) - \bar{\mu}_t(s) \geq \sqrt{\frac{2\ln t}{n_t(s)}}, n_t(s) = t'\right) \qquad (1.8) \\
&= \sum_{t'=1}^{t} \mathbb{P}\left(\mu(s) - \bar{\mu}_t(s) \geq \sqrt{\frac{2\ln t}{n_t(s)}} \middle| n_t(s) = t'\right) \cdot \mathbb{P}\left(n_t(s) = t'\right) \\
&\leq \sum_{t'=1}^{t} \exp\left(-\frac{4t'\ln t}{t'}\right) \cdot \mathbb{P}\left(n_t(s) = t'\right) \qquad (1.9) \\
&= \exp\left(-4\ln t\right) \sum_{t'=1}^{t} \mathbb{P}\left(n_t(s) = t'\right)
\end{aligned}$$

$$\leq \frac{1}{t^4}\,, \tag{1.10}$$

where in (1.8) we use the law of total probability[5]; in (1.9) we use Theorem 1.1, and for (1.10) we use the fact that $\sum_{t'=1}^{t} \mathbb{P}\left(n_t(s) = t'\right) \leq 1$. By symmetry, the same holds for the lower bound of the confidence interval as well. We conclude that, for a given round $t \geq 1$, the true mean of arm s is within the confidence interval with probability at least $1 - 2/t^4$. Thus, the more time passes, the higher the guarantee that the true mean is within the interval.

On the other hand, note that if a certain arm has not been sufficiently explored at a round when t is very large, we would have a large confidence radius, since $\ln t$ would grow while $n_t(s)$ would remain the same. This would increase the overall value of $i_t(s)$ and thus enhance exploration of that arm. The argument for its exploitation is intuitive, as explained earlier.

The UCB1 algorithm [5] selects, in every round, the arm with the highest index, until termination. This strategy proves to be optimal, in that the expected regret incurred by it is logarithmic in T. The authors of [5] prove that

$$R(T) = \left[\sum_{s:\mu(s)<\mu(s^*)} \left(\frac{\ln T}{\Delta_s}\right)\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{s=1}^{N} \Delta_s\right)\,.$$

On the other hand, Lai and Robbins showed that one cannot do better than this, that is, the expected regret of any consistent policy is lower bounded by a constant multiple of $\log T$. In other words, for every suboptimal arm s, we have that $\mathbb{E}[n_t(s)] \geq (\ln t)/D(p^*||p_s)$ asymptotically, where p^* denotes the distribution of s^*, p_s the distribution of s and $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

1.1.4 Contextual bandits

An important extension of the standard MAB is the contextual MAB [10–13], where at the beginning of each round the learner observes side-information, also called the

[5]Since we initially sample each arm once, we do not take into consideration the case when $n_t(s) = 0$.

context, about the arm rewards in that particular round. As the learner tries to maximize its cumulative reward by taking this information into account, its regret is typically measured with respect to an oracle that selects in each round the best arm given the context of that round. Contextual MAB algorithms have been used in a variety of applications ranging from personalized news article recommendation [14] to sequential decision-making in mobile healthcare [15].

For example, in web advertisement, a user's query provides essential information about the nature of the ads the user may potentially click on. Here, the arms can be thought of as user-ad pairs and the context to be used is the actual query. Each time a user query x arrives and an ad s is displayed, the reward collected is a random variable $r(x, s)$ coming from an unknown distribution with mean $\mu(x, s)$. Note that in this case the optimal arm that the oracle chooses (knowing the arms' distribution beforehand) differs from round to round, since the context affects the arms' distributions, and thus, the optimal arm. Assuming a fixed sequence of context arrivals, the regret in T rounds is defined as

$$R(T) = \sum_{t=1}^{T} \mu(x_t, s_t^*) - \sum_{t=1}^{T} \mu(x_t, s_t) \ ,$$

where x_t is the context that arrived in round t and $s_t^* = \text{argmax}_{1 \leq s \leq N} \mu(x_t, s)$.

1.1.5 Combinatorial bandits

Another important extension of the standard MAB is the combinatorial MAB (CMAB), where in each round the learner chooses a subset of base arms, also called the super arm, and obtains a reward that depends on the outcomes of the base arms that are in the chosen super arm [16–21]. This problem is mainly investigated under the semi-bandit feedback setting, where the learner also observes outcomes of the selected base arms. Combinatorial MAB have found applications in slate recommendation [22], crowdsourcing [23] and online influence maximization [24].

In combinatorial bandits, we are in the same conditions with [5], only that the learning agent here selects a subset of the set of base arms called a *super arm*. Subsequently, the agent observes feedback coming from the selected super arm. There are several types of feedback: (i) bandit feedback, where the agent only obtains the reward of the chosen super arm; (ii) semi-bandit feedback, where the learning agent also obtains the stochastic outcomes of the individual base arms constituting the super arm; (iii) cascading feedback [25], where the agent obtains the reward of the chosen super arm and the weights of some of the individual base arms, according to some problem-specific stopping criterion.

The CMAB problem is defined as the classical MAB problem in Section 1.1, where the random variables $r_{s,t}$ have bounded support in $[0, 1]$. A super arm $S = \{s_1, \ldots, s_K\}$ consists of a subset of the set of arms $[N]$, where $K \in \mathbb{N}$. We denote by $\mathcal{S}$ the set of all *feasible*[6] super arms. The stochastic reward vector associated with super arm S is denoted by $r(S) := [r(s_1), \ldots, r(s_K)]^T$. A general assumption in the CMAB literature is that the expected reward function is only a function of the set of arms S and the expected base arm outcome vector $\boldsymbol{\mu} := [\mu(1), \ldots, \mu(N)]^T$. In this case, the regret over T rounds is defined as follows:

$$R(T) = \sum_{t=1}^{T} r_{\boldsymbol{\mu}}(S^*) - \sum_{t=1}^{T} r_{\boldsymbol{\mu}}(S_t) \ ,$$

where we denote by $r_{\boldsymbol{\mu}}(S)$ the expected reward of super arm S; $S^* = \text{argmax}_{S \in \mathcal{S}} \, \mu(S)$ and S_t is the super arm chosen by the algorithm in round t.

The CMAB problem is first considered in [17] under two mild assumptions, namely, monotonicity, which basically presumes that the expected reward of a super arm S_1, with individual base arm whose expected outcomes are all greater than or equal to the expected outcomes of the base arms of a super arm S_2, is greater than or equal to the expected reward of S_2. Furthermore, we have that the expected reward satisfies the bounded smoothness property with respect to the individual expected base arm outcomes. The second assumption gives a measure of continuity of the expected

[6]It will become clear from the context what we will mean by feasible in the following sections.

reward function, relating it to the expected base arm outcomes. Formally, there exists a strictly increasing function $f(\cdot)$, called the *bounded smoothness function*, such that, for any two expectation vectors $\boldsymbol{\mu}$ and $\boldsymbol{\mu'}$ and a given super arm S, we have $|r_{\boldsymbol{\mu}}(S) - r_{\boldsymbol{\mu'}}(S)| \leq f(\Lambda)$, if $\max_{s \in S} |\mu(s) - \mu'(s)| \leq \Lambda$.

Here an (α, β)-approximation oracle is used, which takes as input the expected reward vector $\boldsymbol{\mu}$ and returns the α-optimal super arm with probability at least β, i.e. a super arm S such that $\mathbb{P}\left(\mu(S) \geq \alpha \cdot \mu(S^*)\right) \geq \beta$. Corresponding to this notation, we now define the (α, β)-approximation regret which gives the regret incurred by a policy that uses an (α, β)-approximation oracle with probability at least $1 - \beta$:

$$R_{\alpha,\beta}(T) = \sum_{t=1}^{T} \alpha \cdot \beta \cdot r_{\boldsymbol{\mu}}(S^*) - \sum_{t=1}^{T} r_{\boldsymbol{\mu}}(S_t) \; .$$

A super arm S is called *bad* if $r_{\boldsymbol{\mu}}(S) < \alpha \cdot r_{\boldsymbol{\mu}}(S^*)$. We denote by $\mathcal{S}_B$ the set of all bad super arms. Now, for any base arm s we define

$$\begin{aligned} \Delta^s_{\text{min}} &= \alpha \cdot r_{\boldsymbol{\mu}}(S^*) - \max\{r_{\boldsymbol{\mu}}(S) | S \in \mathcal{S}_B, s \in S\}, \\ \Delta^s_{\text{max}} &= \alpha \cdot r_{\boldsymbol{\mu}}(S^*) - \min\{r_{\boldsymbol{\mu}}(S) | S \in \mathcal{S}_B, s \in S\} \; . \end{aligned}$$

Moreover, let $\Delta_{\text{max}} = \max_{s \in [N]} \Delta^s_{\text{max}}$ and $\Delta_{\text{min}} = \min_{s \in [N]} \Delta^s_{\text{min}}$. The main theorem in [17] states that the (α, β)-approximation regret of CUCB using an (α, β)-approximation oracle is at most

$$R_{\alpha,\beta}(T) \leq \sum_{\substack{s \in [N] \\ \Delta^s_{\text{min}} > 0}} \left(\frac{6(\ln T)\Delta^s_{\text{min}}}{(f^{-1}(\Delta^s_{\text{min}})^2} + \int_{\Delta^s_{\text{min}}}^{\Delta^s_{\text{max}}} \frac{6 \ln T}{(f^{-1}(x))^2} \right) + \left(\frac{\pi^2}{3} + 1 \right) \cdot N \cdot \Delta_{\text{max}} \; ,$$

where $f(\cdot)$ is the bounded smoothness function.

Additionally, another line of the combinatorial bandit research is that of combinatorial bandits with probabilistically triggered arms [18,26,27]. In this setting, base arms' outcomes in a super arm may trigger other base arms to be played. In other words, the learning agent observes semi-bandit feedback from a random superset of the played super arm. The base arm outcome independence holds across different rounds, but not

necessarily across all base arms in a given round.

1.1.6 Volatile (sleeping) bandits

Deviating from the aforementioned works, another strand of literature considers MAB with time-varying arm sets under the names *sleeping* MAB [28, 29] or *volatile* MAB [30], both of which are remnants of *mortal* MAB [31]. In this setting, the learner tries to select the best of the available arms in each round to maximize its cumulative reward. The concept of volatility is quite common in applications that involve sequential decision making. For instance, in online advertising, ads become unavailable after they expire [31]. Similarly, in crowdsourcing, the set of available tasks and workers may change over time [32].

While extending the range of the applications that our methods can accommodate, the volatility assumption also introduces a problem that needs to be solved. Due to volatility, there may be arms that come only a small number of rounds, thereby making accurate estimation of the mean rewards yielded from their selection impossible. For example, if arm s is available for selection only a limited number of times, the accuracy of the estimation of its true mean stops at some point, and this may impede the learning process. In extreme cases this process becomes impossible without further assumptions on the nature of the problem. As we will see, we address this problem by imposing continuity assumptions on the mean function over the base arms.

1.2 Our contribution and related work

In this thesis we study the contextual combinatorial volatile multi-armed bandit (CCV-MAB) problem, which models the repeated interaction between an agent and its dynamically changing environment. In each round t, the agent observes the available base arms and their contexts, selects a subset of the available base arms, which is called a super arm, collects a reward, and observes noisy outcomes of the selected base arms. The goal is to maximize the cumulative reward in a given number of rounds without

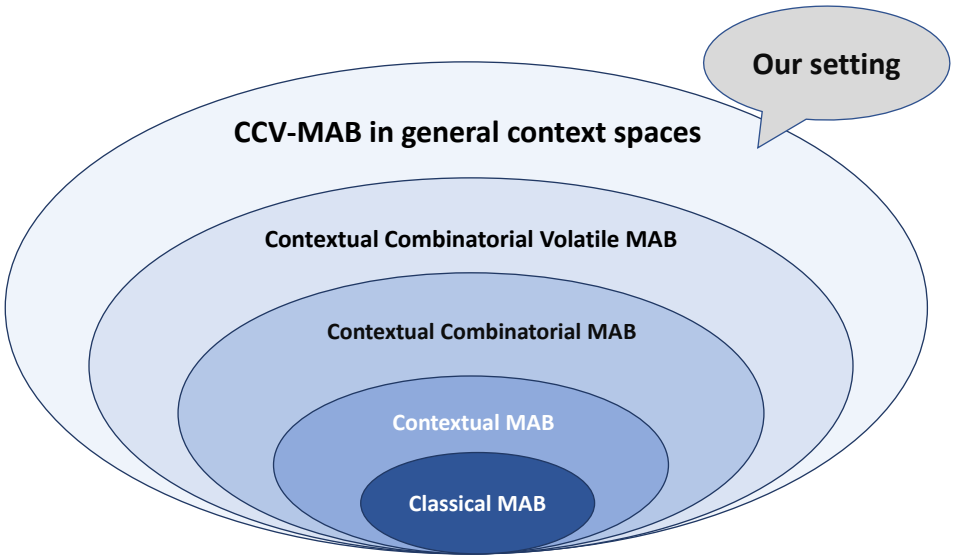

Figure 1.1: The diagram depicts the generality of the CCV-MAB framework comparing to simpler settings.

knowing future context arrivals and the function that maps the contexts to base arm outcomes. Achieving this goal requires careful tuning of exploration and exploitation by adapting decisions in real-time based on the problem structure and past history.

The CCV-MAB model includes the characteristics of all the MAB models mentioned above: (i) base arms are volatile, (ii) expected outcome of a base arm depends on its context, (iii) the learner selects a super arm in each round that consists of a subset of the available base arms, observes outcomes of the selected base arms and receives a reward that depends on these outcomes and (iv) the context space $\mathcal{X}$ is compact. In Table 1.1 we compare the characteristics of our setting with the features of previous work related to ours.

Table 1.1: Comparison with related work. In the first column we refer to the setting we consider in both approaches. The table gives a more detailed exhibition of the generality which our setting accommodates.

Properties	This work	[25]	[33]	[17]	[35]	[13]	[28]
Contextual	Yes	Yes	Yes	No	No	Yes	No
Combinatorial	Yes	Yes	Yes	Yes	No	No	No
Volatile (base) arms	Yes	No	Yes	No	No	Yes	Yes
Arm and/or context space	Infinite	Finite/infinite	Infinite	Finite	Infinite	Infinite	Finite
Adaptive discretization	Yes	No	No	No	Yes	Yes	No
Reward function	General	General	Submodular	General	General	General	General
Feedback type	Semi-bandit	Cascading	Semi-bandit	Semi-bandit	Full-bandit	Full-bandit	Full-bandit
Computation oracle	α	α	$(1-1/e)$	(α,β)	Exact	Exact	Exact

1.2.1 Previous work on CCV-MAB

The Contextual Combinatorial Multi-armed Bandit (CCMAB) problem has been attracting a lot of interest in recent years [33, 34]. In [33] the authors study the CCMAB problem with cascading feedback; that is, the learning agent can observe base arm outcomes of only a prefix of the played super arm. Their model allows for position discounts and a wide range of reward functions. Under monotonicity and Lipschitz continuity of the expected reward, their algorithm, C^3-UCB, incurs $\tilde{O}(\sqrt{KT})$ regret in T rounds, where K is the maximum cardinality of any feasible super arm.

The contextual combinatorial volatile multi-armed bandit (CCV-MAB) has been studied in [33, 36]. In [33] the expected reward is assumed to be submodular, and the expected base arm outcomes are assumed to be Hölder continuous with respect to their corresponding contexts. Their proposed algorithm (CCMAB) addresses the potential volatility of the arms by uniformly discretizing the context space into a predetermined number of hypercubes (which depends on the time horizon T), thus utilizing the outcome similarities between nearby contexts. CCMAB incurs $\tilde{O}(T^{(2\eta+D)/(3\eta+D)})$ regret where η is the Hölder constant and D the context space dimension.

A potential drawback of this approach is the fixed discretization of the context space, which implies limited exploitation of the arms' similarity information. We address this issue by: (i) adaptively discretizing the D-dimensional context space $(\mathcal{X}, \|\cdot\|_2)$ following a tree structure, under some mild structural assumptions on $(\mathcal{X}, \|\cdot\|_2)$ (where $\|\cdot\|_2$ is the Euclidean norm) and Lipschitz continuity assumption both of the expected reward with respect to the expected base arm outcomes and the latter

with respect to their associated contexts; (ii) imposing a GP structure on the expected base arm outcomes and thus using the GP posterior properties to give estimates of the true means of all arms. Below we explain these two ways in general.

1.2.2 Two approaches to the high cardinality problem

In our setting, the context space $\mathcal{X}$ is assumed to be compact, and thus, there may be infinitely many contexts associated with the base arm set. This introduces a problem: an infinite context space directly affects exploration. Blindly exploring infinitely many arms without any means of navigation is like trying to find a needle in a haystack. Therefore, mild continuity assumptions on the expected base arm outcomes and the expected reward function (expected reward from the selected super arm) are necessary. Using these two assumptions, we use the so called adaptive discretization technique. The context space is first partitioned into two identical regions and statistics of randomly selected arms with contexts in these regions are stored for each region. As the confidence radii of these regions become smaller and smaller, we decide to partition them into more and more refined regions, based on a rule that relates this radius with the region's geometric radius[7]. A region therefore contains similar contexts, that is, contexts that yield similar outcomes. Theoretically, a region converges to a single context in limit, after an infinite number of partitions.

Adaptive discretization has been utilized in the GP bandit setting in [37] for black-box function optimization. Under the same structural assumptions of the arm space, $(\mathcal{X}, d)$ as in [36] (where d is a general metric associated with $\mathcal{X}$), and under some Hölder continuity assumptions on their covariance function, they propose a tree-based algorithm whose aim is to maximize the cumulative reward given a fixed budget of samples. They provide both near-optimality dimension type and information type regret bounds. It is worth mentioning that when the context space is very large, and the arm set is time-varying, adaptive discretization has proved to be an efficient technique, yielding optimal regret bounds [35].

[7]The details of the expansion rule are given in Section 3.1.

However, if we further assume that the number of arms that come in a round is finite, imposing a GP prior[8] on the base arm outcomes removes both the need for adaptive discretization and Lipschitz continuity[9] of the expected base arm outcomes. The posterior's smoothness encodes enough information to estimate with high probability the outcomes of any available base arm, regardless of its history of arrivals.

This brings us to our second method, which, as we will see, achieves better results, both theoretically and experimentally. While rigid parametric assumptions such as linear realizability [21] lead to appealing theoretical analysis, they may not be accurate in modeling complex outcome functions. On the other hand, the smoothness encoded by the GPs put sufficient structure on the problem such that it is efficiently learnable while still capturing a wide range of functions of practical interest [38]. This motivates us to model the base arm outcome function as a sample from a GP with a known kernel.

GP-based bandit algorithms have been commonly used to address the exploration-exploitation dilemma since the pioneering work of [39]. The closed-form expressions for their posterior mean and variance give rise to high probability confidence bounds on the expected value of the function to be learned, thereby inducing an optimal balance between exploration and exploitation. An interesting feature that grants GP-UCB superiority to its frequentist counterpart is its ability to give high probability estimates to actions that have not been played before due to the posterior's smoothness. The frequentist-type UCB algorithms need explicit smoothness assumptions to make this possible. In a bandit setting with time-varying action sets, this feature proves extremely useful; it inherently encodes the necessary structural properties needed for learning.

1.2.3 Our contribution

Now we briefly explain our first approach to the CCV-MAB problem, namely, adaptive discretization. Here, base arms are defined with respect to their corresponding contexts, which lie in a compact metric space endowed with some "nice" structural

[8]Note that this is, in itself, another assumption on the nature of the expected base arm outcomes.
[9]Note that the GP inherently assumes that the latent function satisfies smoothness conditions.

properties[10]. In every round, the learning agent observes a time-varying set of available base arms (and their corresponding contexts) and consequently selects a super arm. Selection of the arms yields stochastic feedback coming from a distribution unknown to the learner. Following the super arm selection, semi-bandit feedback is observed. The learner's goal is to maximize the cumulative reward over all rounds up to a fixed horizon.

Two critical problems that instantly introduce themselves in this setting, making the classical bandit approaches unsuitable, are volatility and the context space's large cardinality. Under Lipschitz continuity assumption of the expected reward function (expected base arm outcomes) with respect to expected base arm outcomes (contexts), we propose *Adaptive Contextual Combinatorial Upper Confidence Bound* ACC-UCB, an online learning algorithm that partitions $\mathcal{X}$ into confidence regions, centered at contexts called nodes, following a tree structure. Every active leaf node is associated with an index, which is simultaneously a high probability upper bound on the base arms' expected outcomes in the associated region and the variation of the expected outcome values of different arms in the region. When the uncertainty of the expected outcomes associated with the contexts in a region is lower than the region's diameter, ACC-UCB decides to partition that region into smaller subregions further. In this way, the similarity information between nearby contexts is fully utilized. ACC-UCB incurs $\tilde{O}(T^{(\bar{D}+1)/(\bar{D}+2)+\epsilon})$ regret for any $\epsilon > 0$, where $\bar{D}$ is the approximate optimality dimension of the context space $\mathcal{X}$, tailored to capture both the benignness of the base arm arrivals and the structure of the expected reward under a combinatorial setup. In general, $\bar{D} \leq D$, which implies an improvement of the bound rates to the previous previous state-of-the-art [33]. Furthermore, empirical superiority to the previous state-of-the-art is also demonstrated via experimental results. Our model and algorithm can be applied to solve dynamic resource allocation problems ranging from crowdsourcing to online advertising to multi-user channel allocation [40].

Second, we propose another learning algorithm called *Optimistic Combinatorial Learning and Optimization with Kernel Upper Confidence Bounds* (O'CLOK-UCB),

[10]See Definition 2.1.

tailored to solve the CCV-MAB problem over compact context spaces under the assumption that the expected base arm outcomes are samples from a GP and the expected reward is Lipschitz for the expected base arm outcomes. O'CLOK-UCB incurs $\tilde{O}\left(K\sqrt{T\overline{\gamma}_T}\right)$ regret in T rounds in the general case when the feasible action cardinalities are bounded variables, where $\overline{\gamma}_T$ is the maximum information gain over the feasible actions with cardinalities upper bounded by $K \in \mathbb{N}$.

The theoretical analysis is an important contribution of this work. Simultaneously taking into consideration several bandit features introduces new theoretical difficulties, which we solve as follows:

- the semi-bandit feedback necessitates a batch update of all the base arm indices at once after all K arms are selected. This, in return, requires a new way of upper-bounding the expected regret, substantially different from previous related work [39]. We prove a new auxiliary result which helps to achieve our bounds;
- furthermore, a new notion of maximum information gain is needed for the CCV-MAB framework in order to obtain tight bounds. We introduce $\bar{\gamma}_T$ which takes into consideration all of this framework's traits and provide regret bounds that depend on it;
- in order to better understand the nature of this new notion we bound it by the classical notion of maximum information gain and provide an explicit role that K plays in these bounds, given in Theorem 4.3. This also allows us to give explicit regret bounds and show the role of both K and T in them.

In addition to capturing a large class of bandit instances, our general setup captures the scenario that the base arm outcomes depend on each other which makes for tighter bounds since it provides more information between arms. We prove lower and upper bounds on $\overline{\gamma}_T$ in terms of the classical notion of maximum information gain γ_T in the case when the cardinality of any feasible action is a constant K, in which case, O'CLOK-UCB incurs $\tilde{O}\left(K\sqrt{T\gamma_T}\right)$ regret. Moreover, we derive regret bounds that show explicit dependence on both K and T for the commonly used kernels and give an exact comparison between our bounds and those of previous work, thereby exhibiting a substantial improvement on theoretical analysis. We perform semi-synthetic

crowdsourcing simulations on a real-world dataset and compare our algorithm with the previous state-of-the-art. Moreover, we improve the time complexity of O'CLOK-UCB by using the sparse approximation method, where we only use a subset of the past history of selected contexts to update the posterior. We experimentally show that this method outperforms the previous state-of-the-art UCB-based algorithms for this setting, including ACC-UCB.

1.3 Applications of the CCV-MAB model

The CCV-MAB framework can be used to model many practical scenarios. Features such as time-variant arm set availability (volatility), multiple arms selection and the presence of side information (context) are very realistic when you think about scenarios such as recommender systems [34], multi-user small base station association [40] and adaptively controlled trials [41].

1.3.1 Recommender systems

In recommender systems we usually have a list of items (eg. movies, books, sites, etc.) which have to be recommended to different users. User profiles are usually accessible. In the movie recommendation example, the past history of the user, his preferences, his age and so on, can be regarded as side information about the user. On the other hand, the movie parameters, such as genre, date of production, featuring stars, etc., can be regarded as side information about the movie. Now, each movie-user pair is regarded as an arm, and all the parametrized aforementioned side information comprise the context vector. The context space is thus a vector space[11] and some of the context parameters can be mapped into some interval, and may potentially take any value within it. Thus the compactness assumption of the context space is advantageous to learning. In each round the streaming platform may recommend multiple movies, a feature that is captured by the combinatorial nature of the CCV-MAB problem. Moreover, some movies

[11] Note that some of the dimensions of the context vector can be discrete values.

may not be available for streaming in some countries at certain periods of time, which is represented by our volatility assumption. Thus the CCV-MAB is a natural model for this scenario.

Formally, in each round $t \geq 1$, the streaming platform recommends K movies $m_{t,1}, \ldots, m_{t,K}$ to user u_t. Therefore the selected super arm is formally defined as $S_t = \{(m_{t,1}, u_t), \ldots, (m_{t,K}, u_t)\}$, and each individual base arm $s_t := (m_{t,j}, u_t)$ has an associated context vector $\boldsymbol{x}_{t,s_t}$. Afterwards, semi-bandit feedback is observed which in this case may be 0 if the user didn't watch the movie within a certain time period, or 1 otherwise.

1.3.2 Multi-user small base station association

In [40] the authors solve the problem of assigning small base stations (SBSs) to multiple users in order to maximize throughput. We have M SBSs and each SBS m can serve at most q_m users. The allocated slot for a user in an SBS is called a channel (SBSC). Thus, a user-SBSC pair is regarded as an arm. Note that, in a certain area, there may be users on the move, and thus their availability is time-variant with regard to the SBSs in that area. This is captured by the volatility feature of CCV-MAB. The available users at time t are denoted by the set $\mathcal{N}_t$ and the channels of all SBSs are represented by the set $\mathcal{C}$. Thus, the set of available arms at time t are defined by such pairs. Formally, we have $\mathcal{A}_t := \{(c, n) : c \in \mathcal{C}, n \in \mathcal{N}_t\}$.

At each round $t \geq 1$, a central processor (CP) associates users to SBSCs and, for each association $a = (c, n)$, a reward (which in this case is the throughput) is observed. In the SBS case, the reward is affected by the transmission frequency and the distance between the user and the station. Thus, we have a two-dimensional context x which helps the learning process. Therefore, for a given arm a, the reward observed at round t is denoted by $g(x_{t,a})$, where $x_{t,a}$ is the context of arm a at time t. On the other hand, in each round the CP has to associate all the available users[12] to the SBSCs and then observes throughput of all of them. Thus, we can see this as semi-bandit feedback.

[12]Here there is a constraint to be considered. Each user can be associated to only one SBSC.

A subset of the set of potential associations is called a super association (which can be thought of as a super arm). The last thing that completes the picture is the fact that the context dimensions (frequency and distance) are normalized using a realistic maximum value for both of them, and thus, they potentially take any value between 0 and 1. This accounts for the compact context space. All of the features above call for a CCV-MAB application to this problem.

1.3.3 Adaptively controlled trials

Randomized controlled trials are a statistical method that helps decide whether a new treatment is better than the old one or not by randomly dividing the recruited patients into treatment and control groups. As shown in [41], this method can be highly suboptimal. Another way to recruit the patients is by cohorts, not all at once, and then use the data gathered by the previous cohorts to the future ones. This problem can be modeled as a sequential decision making one. The process may be repeated many times in order to get better results.

Formally, there are N patients to be recruited in K steps. At step k, a total of M_k patients can be recruited. This feature is captured by the combinatorial nature of CCV-MAB. The fact that, up until the first K rounds, the patients selected in a group will not be available for selection, is captured by volatility. Furthermore, each patient-treatment pair is associated with side information such as patient medical history, treatment dosage, time of intake, etc., which can be seen as context. Thus, we are in conditions of a CCV-MAB problem.

1.4 Organization

The rest of the thesis is organized as follows. In Chapter 2 we formalize the CCV-MAB problem by introducing the notation to be used, the necessary assumptions to be made for both methods, and some technical definitions. Some examples of the CCV-MAB problem conclude the chapter. In Chapter 3 we introduce our first method, ACC-UCB,

and then analyze its regret. The chapter concludes with experimental results. The same is the case in Chapter 4 where we deal with O'CLOK-UCB. Chapter 5 concludes the thesis.

Chapter 2

CCV-MAB problem formulation

In this chapter we will formally introduce the CCV-MAB setting. In the first section we start by defining the necessary notation and assumptions required. The second section consists of the problem definition. In the third section we give some examples of problems where the CCV-MAB model can be applied. Sections 2.4 and 2.5 give the technical definitions related to the adaptive discretization and the Gaussian process methods, respectively.

2.1 General notation

We first introduce the general notation to be used throughout the thesis. Let us fix a positive integer $K \geq 1$ and a D-dimensional vector space $\mathcal{X}$. We write $[K] = \{1, 2, \ldots, K\}$. Vectors are denoted by bold lowercase letters. Furthermore, given $N \in \mathbb{N}$, $\boldsymbol{I}_N$ denotes the identity matrix in $\mathbb{R}^{N \times N}$ and $\mathbb{I}(\cdot)$ denotes the indicator function.

2.1.1 Base arms and their outcomes

Let $(\mathcal{X}, d)$ be a metric space. Our setup involves base arms that are defined by their contexts, x, belonging to the context set $\mathcal{X}$. Let $r(x)$ represent the random outcome generated by the base arm with context x. We assume that there exists a function $f : \mathcal{X} \to \mathbb{R}$, such that we have $r(x) = f(x)+\eta$, for every $x \in \mathcal{X}$, where $\eta \sim \mathcal{N}(0, \sigma^2)$. We assume that η is independent across base arms.

2.1.2 Super arms and their rewards

A super arm S is a feasible subset of the base arm set and it is defined by the contexts of its base arms. Consider a super arm S associated with the context tuple $\boldsymbol{x} = [x_1, \ldots, x_{|S|}]^T$ such that $x_m \in \mathcal{X}$, $\forall m \leq |S|$. The corresponding outcome and expected outcome vectors (the latter is also called the expectation vector) are denoted by $r(\boldsymbol{x}) = [r(x_1), \ldots, r(x_{|S|})]^T$ and $f(\boldsymbol{x}) = [f(x_1), \ldots, f(x_{|S|})]^T$.

We assume that the reward received from playing this super arm is a non-negative random variable denoted by $U(r(\boldsymbol{x}))$. Moreover, we assume that the expected reward of playing any super arm is a function only of the set of arms and the mean vector, and we define $u(f(\boldsymbol{x})) = \mathbb{E}[U(r(\boldsymbol{x}))|f]$. Again, this assumption is standard in combinatorial bandits [17]. In addition, we impose the following mild assumptions on u, which allow for a very large class of functions to fit our model such as multi-armed bandit with multiple plays, maximum weighted bipartate matching and probabilistic maximum coverage [17].

Assumption 1. *(Monotonicity) Let S be a feasible super arm. For any $\boldsymbol{f} = [f_1, \ldots, f_{|S|}]^T \in \mathbb{R}^{|S|}$ and $\boldsymbol{f'} = [f'_1, \ldots, f'_{|S|}]^T \in \mathbb{R}^{|S|}$, if $f_m \leq f'_m$, $\forall m \leq |S|$, then $u(\boldsymbol{f}) \leq u(\boldsymbol{f'})$.*

Assumption 2. *(Lipschitz continuity of the expected reward in expected outcomes) $\exists B > 0$ such that for any feasible super arm S, $\boldsymbol{f} = [f_1, \ldots, f_{|S|}]^T \in \mathbb{R}^{|S|}$ and $\boldsymbol{f'} = [f'_1, \ldots, f'_{|S|}]^T \in \mathbb{R}^{|S|}$, we have $|u(\boldsymbol{f}) - u(\boldsymbol{f'})| \leq B \sum_{i=1}^{|S|} |f_i - f'_i|$.*

Assumption 1 states that the expected reward is monotonically non-decreasing with

respect to the expected outcome vector. Assumption 2 implies that the expected reward varies smoothly as a function of expected base arm outcomes.

2.2 The optimization and learning problems

We consider a sequential decision-making problem with volatile base arms that proceeds over T rounds indexed by $t \in [T]$. The agent knows u perfectly, but does not know f beforehand. In each round t, M_t base arms indexed by the set $\mathcal{M}_t = [M_t]$ are available. We assume $\max_{t \geq 1} M_t \leq M$, for some integer M. The context of base arm $m \in \mathcal{M}_t$ is represented by $x_{t,m} \in \mathcal{X}$. We denote by $\mathcal{X}_t = \{x_{t,m}\}_{m \in \mathcal{M}_t}$ the set of available contexts and by $\boldsymbol{f}_t = [f(x_{t,m})]_{m \in \mathcal{M}_t}^T$ the vector of expected outcomes of the available base arms in round t. We denote by $\mathcal{S}_t$ the set of feasible super arms in round t and $\mathcal{S} = \cup_{t \geq 1} \mathcal{S}_t$ the overall feasible set of super arms. Note that $\mathcal{S}_t$ depends on the structure of the optimization problem which is known by the learner, and if $S \in \mathcal{S}_t$, then $S \subseteq \mathcal{M}_t$. Furthermore, we assume that the budget (maximum number of base arms in a super arm) does not exceed some fixed integer $K \in \mathbb{N}$, that is, for any $S \in \mathcal{S}$, we have $|S| \leq K$. At the beginning of round t, the agent first observes $\mathcal{M}_t$ and $\mathcal{X}_t$. Then, it selects a super arm S_t from $\mathcal{S}_t$.

2.2.1 The optimization problem

For a moment, consider the hypothetical situation where f were known beforehand. Then the agent would have selected an optimal super arm given by $S_t^* \in \text{argmax}_{S \in \mathcal{S}_t} u(f(\boldsymbol{x}_{t,S}))$, and will obtain an expected reward of $\text{opt}(\boldsymbol{f}_t) = \max_{S \in \mathcal{S}_t} u(f(\boldsymbol{x}_{t,S}))$. However, it is known that in general combinatorial optimization is NP-hard [42], and thus, efficient computation of S_t^* is not possible even if f were known. Fortunately, for many combinatorial optimization problems of interest, there exist computationally efficient approximation oracles. In the previous chapter we introduced the notion of an (α, β)-approximation oracle. There is also another notion which

takes into consideration that the oracle may only return a fraction of the optimal solution with certainty. If this is the case, we have what is called an α-approximation oracle. In this thesis, we assume that the agent obtains S_t from an α-approximation oracle, which, when given as input $\boldsymbol{f}_t$ and the specific structure of the particular optimization problem, returns a super arm Oracle(f_t)[1] such that $u(f(\boldsymbol{x}_{t,\text{Oracle}(f_t)})) \geq \alpha \times \text{opt}(\boldsymbol{f}_t)$.

2.2.2 The learning problem

Since the agent does not know f in our case, it calls the approximation oracle in each round t with an M_t-dimensional parameter vector $\boldsymbol{\theta}_t$ to get $S_t = \text{Oracle}(\boldsymbol{\theta}_t)$. Here, $\boldsymbol{\theta}_t$ is an approximation to $\boldsymbol{f}_t$ calculated based on the history of observations. Through our analyses, we will show that our learning algorithms use upper confidence bounds (UCBs) as $\boldsymbol{\theta}_t$, based on a frequentist and Bayesian approach. It is important to note here that S_t is an α-optimal solution under $\boldsymbol{\theta}_t$ but not necessarily under $\boldsymbol{f}_t$.

At the end of round t, the agent collects the reward $U(r(\boldsymbol{x}_{t,S_t}))$ where $\boldsymbol{x}_{t,S_t} = [x_{t,s_{t,1}}, \ldots, x_{t,s_{t,|S_t|}}]^T$ is the set of context vectors associated with super arm S_t and $r(\boldsymbol{x}_{t,S_t}) = [r(x_{t,s_{t,1}}), \ldots, r(x_{t,s_{t,|S_t|}})]^T$ is the outcome vector of super arm S_t. It also observes $r(\boldsymbol{x}_{t,S_t})$ as a part of the semi-bandit feedback. The goal of the agent is to maximize its expected cumulative reward in the long run.

To measure the loss of the agent in this setting by round T for a given sequence of base arm availability $\{\mathcal{X}_t\}_{t=1}^T$, we use the standard notion of α-approximation regret (referred to as the regret hereafter), which is given as

$$R_\alpha(T) = \alpha \sum_{t=1}^{T} \text{opt}(\boldsymbol{f}_t) - \sum_{t=1}^{T} u(f(\boldsymbol{x}_{t,S_t})) \ .$$

[1]Note that Oracle also takes as input the feasible set $\mathcal{S}_t$, but this is omitted from the notation for brevity.

2.3 Additional example applications of CCV-MAB

In this section, we detail several applications of CCV-MAB.

Multi-armed bandit with volatile arms and multiple plays. In this problem, $U(r(\boldsymbol{x}_{t,S})) = \sum_{m \in S} r(x_{t,m})$ and $u(f(\boldsymbol{x}_{t,S})) = \sum_{m \in S} f(x_{t,m})$. S_t^* simply consists of K base arms in $\mathcal{X}_t$ with the highest expected outcomes. Hence, instead of an approximation oracle, an exact oracle ($\alpha = 1$) with polynomial computation time can be used.

Dynamic maximum weighted bipartate matching. Let $G_t = (L_t, R, E_t)$ represent the bipartate graph in round t, where L_t and R are the set of nodes that form the parts of the graph such that $|L_t| \geq K$ and $|R| = K$ (here, K represents the budget), and E_t is the set of edges. Here, each edge $(i, j) \in E_t$, such that $i \in L_t$ and $j \in R$, represents a base arm. Weight of edge $(i, j) \in E_t$ is given by $f(x_{t,(i,j)})$. In this problem, $\mathcal{S}_t$ corresponds to the set of K-element matchings of G_t. Hence, given $S \in \mathcal{S}_t$, $U(r(\boldsymbol{x}_{t,S})) = \sum_{m \in S} r(x_{t,m})$ and $u(f(\boldsymbol{x}_{t,S})) = \sum_{m \in S} f(x_{t,m})$. S_t^* can be computed by the Hungarian algorithm or its variants [43] in polynomial computation time, and hence, $\alpha = 1$. This problem can model dynamic assignment of R resources to $|L_t|$ available users in round t, where each resource can be assigned to at most one user. One example application is multi-user multi-channel communication [21].

Dynamic probabilistic maximum coverage. Let $G_t = (L_t, R_t, E_t)$ represent the bipartate graph in round t, where L_t and R_t are the set of nodes that form the parts of the graph, and E_t is the set of edges such that $|L_t| \geq K$. Each edge $(i, j) \in E_t$ has a context dependent activation probability given as $f(x_{t,(i,j)})$, i.e., when a node $i \in U_t$ is chosen, it activates its neighbor j with probability $f(x_{t,(i,j)})$ independent of the other neighbors of j. Here, the goal is to choose a subset of K nodes in U_t that maximizes the expected number of activated nodes in V_t. While it is known that this problem is NP-hard [18], there exists a deterministic $\alpha = (1 - 1/e)$-approximation oracle [44].

2.4 The approximate optimality dimension

We first give the definition of a well-behaved metric space.

Definition 2.1. *(Well-behaved metric space [35]) A compact metric space $(\mathcal{X}, d)$ is said to be well-behaved if there exists a sequence of subsets $(\mathcal{X}_h)_{h\geq 0}$ of $\mathcal{X}$ satisfying the following properties:*

1. *Given $N \in \mathbb{N}$, each subset $\mathcal{X}_h$ has N^h elements, i.e. $\mathcal{X}_h = \{\chi_{h,i},\ 1 \leq i \leq N^h\}$ and to each element $\chi_{h,i}$ is associated a cell $X_{h,i} = \{x \in \mathcal{X} :\ d(x, \chi_{h,i}) \leq d(x, \chi_{h,j}),\ \forall j \neq i\}$.*

2. *For all $h \geq 0$ and $1 \leq i \leq N^h$, we have: $X_{h,i} = \cup_{j=N(i-1)+1}^{Ni} X_{h+1,j}$. The nodes $\chi_{h+1,j}$ for $N(i-1)+1 \leq j \leq Ni$ are called the children of $\chi_{h,i}$, which in turn is referred to as the parent.*

3. *We assume that the cells have geometrically decaying radii, i.e. there exists $0 < \rho < 1$ and $0 < v_2 \leq 1 \leq v_1$ such that we have $B(\chi_{h,i},\ v_2\rho^h/2) \subseteq X_{h,i} \subseteq B(\chi_{h,i},\ v_1\rho^h/2)$, where $B(x, r)$ denotes a closed ball centered at x with radius r. Note that we have $v_2\rho^h \leq \text{diam}(X_{h,i}) \leq v_1\rho^h$, where $\text{diam}(X_{h,i}) := \sup_{x,y\in X_{h,i}} d(x, y)$.*

The first property implies that for every $h \geq 0$ the cells $X_{h,i},\ 1 \leq i \leq N^h$ partition $\mathcal{X}$. This can be observed trivially by reductio ad absurdum. The second property intuitively means that as h grows, we get a more refined sequence of partitions. The third property implies that the nodes $\chi_{h,i}$ are evenly spread out in the space.

The regret bounds of ACC-UCB depend on the notion of approximate optimality dimension, which relates to the dimensions of sets of optimistic contexts that yield approximately optimal expected rewards. Let us now define this dimension, tailored to our combinatorial setting, which is inspired by definitions of the near optimality dimension given in [35, 37, 45].

Definition 2.2. *(The approximate optimality dimension)*

- *A subset $\mathcal{X}_2$ of $\mathcal{X}$ is called r-separated if for any $x_1, x_2 \in \mathcal{X}_2$ such that $x_1 \neq x_2$, we have $d(x_1, x_2) \geq r$. The cardinality of the largest such set is called the **r-packing** number of $\mathcal{X}$ with respect to d, and is denoted by $M(\mathcal{X}, d, r)$. Equivalently, the r-packing number of $\mathcal{X}$ is the maximum number of disjoint d-balls of radius r that are contained in $\mathcal{X}$.*

- *Let $\mathcal{Z} = \{\boldsymbol{x} \in \mathcal{X}^{|S|} : S \in \mathcal{S}\}$. For any $\kappa > 0$, $r > 0$, $t > 0$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$, we define the set*

$$\mathcal{X}^{\kappa}_{g(r)} := \{x \in \mathcal{X} : \kappa - u(f(\boldsymbol{x})) \leq g(r)\,, \text{for some } \boldsymbol{x} \in \mathcal{Z} \text{ such that } x \in \boldsymbol{x}\}$$

*to be an $(g(r), u, \kappa)$-**optimal** set. Let $M(\mathcal{X}^{\kappa}_{g(r)}, d, r)$ be its r-packing number. We define the (g, u, κ)-**optimality dimension** $D^u(g, \kappa)$ associated with $\mathcal{X}^{\kappa}_{g(r)}$ and u as follows:*

$$D^u(g, \kappa) = \max\left\{0, \limsup_{r \to 0} \frac{\log(M(\mathcal{X}^{\kappa}_{g(r)} d, r))}{\log(r^{-1})}\right\}$$

Note that if $\boldsymbol{x} \in \mathcal{Z}$ is such that $\kappa - u(f(\boldsymbol{x})) \leq g(r)$, then all elements of $\boldsymbol{x}$ will be in $\mathcal{X}^{\kappa}_{g(r)}$. Use of the κ-optimality dimension allows us to bound the regret in the volatile setting in a way that the time order of the regret depends on $D^u(g, \kappa)$ which is in most of the cases strictly smaller than D as opposed to the prior work [33] which has bounds that depend on D. For instance, we can let $u^*_{\min} = \min_{t \in [T]} u(f(\boldsymbol{x}^t_{S^{*t}}))$ and $\kappa = \alpha u^*_{\min}$ to obtain a worst-case approximate optimality dimension $\bar{D} = D^u(g, \alpha u^*_{\min})$.

Remark 1. *$\bar{D}$ is obviously less than or equal to D since we are narrowing the space to specific subspaces. We illustrate this in the example below, which is a modified version of Example 3 in [35]. We thus exploit the nature of the reward function u. If we remove the volatility assumption from the problem setting and assume that the learner can choose only one arm in a given round, $\bar{D}$ will be the near optimality dimension of the context space (where we let κ be the optimal expected reward), thus recovering the notion as introduced in previous works. In Section 3.2.2 we explain ways to construct more optimistic regret bounds using the approximate optimality dimension. Moreover, a case for which $\bar{D} < D$ when $\alpha < 1$ is given below, in Example 1.*

There is an important result that relates the approximate optimality dimension with the packing number of the context space. We will make use of it proving the regret bounds of ACC-UCB and in Example 1.

Lemma 2.1. *Fix $\kappa > 0$. Let $\bar{D} = D^u(g, \kappa)$ and $g(r) = cr$ for a given $c > 0$. Fix $D_1 > \bar{D}$. Then, there exists a constant Q, such that for all $r \leq v_2$ we have $M(\mathcal{X}_{cr}^{\kappa}, d, r) \leq Qr^{-D_1}$.*

Proof. By Definition 2.2 we have that

$$\limsup_{r \to 0} \frac{\log(M(\mathcal{X}_{cr}^{\kappa}, d, r))}{\log(r^{-1})} \leq \bar{D}$$

therefore, since $D_1 > \bar{D}$, there exists $r(D_1)$ such that for all $r \leq r(D_1)$ we have

$$\frac{\log(M(\mathcal{X}_{cr}^{\kappa}, d, r))}{\log(r^{-1})} \leq D_1$$

and thus, for all $r \leq r(D_1)$, we have

$$M(\mathcal{X}_{cr}^{\kappa}, d, r) \leq r^{-D_1} .$$

If $r(D_1) \geq v_2$, then we let $Q = 1$ and we are done. If $r(D_1) < v_2$, then we have to show that the result holds for all $r(D_1) \leq r \leq v_2$. First, by the definition of the r-packing number, we have that

$$M(\mathcal{X}_{cr}^{\kappa}, d, r) \leq M(\mathcal{X}, d, r(D_1))$$

since $|\mathcal{X}_{cr}^{\kappa}| \leq |\mathcal{X}|$ and $r \geq r(D_1)$. Thus, for all $r \in [r(D_1), v_2]$, we can say that

$$M(\mathcal{X}_{cr}^{\kappa}, d, r) \leq \frac{v_2^{D_1}}{r^{D_1}} M(\mathcal{X}, d, r(D_1)) \leq Qr^{-D_1}$$

for $Q = \max\{1, M(\mathcal{X}, d, r(D_1)) \cdot v_2^{D_1}\}$. ■

The following example illustrates a case when the approximate optimality dimension is indeed strictly less than the space dimension.

Example 1. *Let $(\mathcal{X}, d) = ([0,1]^D, \|\cdot\|_2)$ and $K = 1$. Let us define $u(f(x)) = 1 - \|x\|_2^a$*

for some $a > 1$. Fix $c > 0$, let $g(r) = cr$ and $\kappa = u^*_{\min}$ (exact oracle). Let $\tilde{t} = \operatorname{argmin}_{t\in[T]} u(f(x^{*t}))$, where $x^{*t} := \operatorname{argmax}_{x\in\mathcal{X}^t} u(f(x))$. Denote $x^* = x^{*\tilde{t}}$ and note that in this case $\kappa = u(f(x^*))$. When $\|x^*\|_2 = 0$, we have $\bar{D} \leq (1 - 1/a)D < D$.

Proof. We have

$$\begin{aligned}
\mathcal{X}^{\kappa}_{c\epsilon} &= \{x \in \mathcal{X} : \kappa - u(f(x)) \leq c\epsilon\} \\
&= \{x \in \mathcal{X} : (1 - \|x^*\|_2^a) - (1 - \|x\|_2^a) \leq c\epsilon\} \\
&= \{x \in \mathcal{X} : \|x\|_2^a \leq c\epsilon\} \\
&= \{x \in \mathcal{X} : \|0 - x\|_2^a \leq ((c\epsilon)^{1/a})^a\} \\
&= \{x \in \mathcal{X} : \|0 - x\|_2 \leq (c\epsilon)^{1/a}\} .
\end{aligned}$$

This set is an $\|\cdot\|_2$-ball with center at 0 and radius $(c\epsilon)^{1/a}$. What we want to know is, for a fixed ϵ, how many disjoint $\|\cdot\|_2$-balls of radius ϵ can fit inside of it (so that we can estimate its ϵ-packing number). Thus, we have

$$M(\mathcal{X}^{\kappa}_{c\epsilon}, \|\cdot\|_2, r) \leq \left(\frac{(c\epsilon)^{1/a}}{\epsilon^1}\right)^D = c^{D/a}(\epsilon^{-1})^{(1-1/a)D} .$$

This together with Lemma 2.1 implies that $\bar{D} \leq (1 - 1/a)D$ for $a > 1$. ■

2.5 Gaussian processes and information gain

Next we introduce some more definitions that will be used in the second approach to the CCV-MAB problem. We also define a new notion of maximum information gain $\bar{\gamma}_T$, which accommodates the CCV-MAB features simultaneously. The regret bounds of O'CLOK-UCB depend on $\bar{\gamma}_T$.

2.5.1 Putting structure on base arm outcomes via GPs

Since the agent does not have any control over M_t and $\mathcal{X}_t$, expected outcomes of available base arms can vary greatly over the rounds. Since how good the learner performs depends on how well it learns the unknown f, we need to impose regularity conditions on f. This regularity assumption may also be formalized by modeling f as a sample from a Gaussian process, which is defined below. In Chapter 4, we will solve the CCV-MAB problem by assuming that f is a sample from a GP.

Definition 2.3. *A Gaussian Process with index set $\mathcal{X}$ is a collection $(f(x))_{x\in\mathcal{X}}$ of random variables which satisfies the property that $(f(x_1),\ldots,f(x_n))$ is a Gaussian random vector for all $\{x_1,\ldots,x_n\} \in \mathcal{X}$ and $n \in \mathbb{N}$. The probability law of a GP $(f(x))_{x\in\mathcal{X}}$ is uniquely specified by it mean function $x \mapsto \mu(x) = \mathbb{E}[f(x)]$ and its covariance function $(x_1,x_2) \mapsto k(x_1,x_2) = \mathbb{E}[f(x_1)-\mu(x_1)][f(x_2)-\mu(x_2)]$.*

2.5.2 The posterior distribution of outcomes

In this subsection, we recall the closed form expressions for the posterior distribution of GP-sampled functions, given a set of observations. Fix $N \in \mathbb{N}$. Assuming for now that our function is a sample from a GP, consider a finite sequence $\tilde{\boldsymbol{x}}_{[N]} = [\tilde{x}_1,\ldots,\tilde{x}_N]^T$ of contexts with the corresponding vector $\boldsymbol{r}_{[N]} := r_{[N]}(\tilde{\boldsymbol{x}}_{[N]}) = [r(\tilde{x}_1),\ldots,r(\tilde{x}_N)]^T$ of outcomes and corresponding vector $\boldsymbol{f}_{[N]} = [f(\tilde{x}_1),\ldots,f(\tilde{x}_N)]^T$ of expected outcomes. For every $n \leq N$, we have $r(\tilde{x}_n) = f(\tilde{x}_n) + \eta_n$, where η_n is the noise that corresponds to this particular outcome.

The posterior distribution of f given $r_{[N]}$ is that of a GP with mean function μ_N and covariance function k_N given by

$$\mu_N(\tilde{x}) = (k_{[N]}(\tilde{x}))^T(\boldsymbol{K}_{[N]} + \sigma^2\boldsymbol{I}_N)^{-1}r_{[N]}, \tag{2.1}$$

$$k_N(\tilde{x},\tilde{x}') = k(\tilde{x},\tilde{x}') - (k_{[N]}(\tilde{x}))^T(\boldsymbol{K}_{[N]} + \sigma^2\boldsymbol{I}_N)^{-1}k_{[N]}(\tilde{x}') \tag{2.2}$$

$$\sigma_N^2(\tilde{x}) = k_N(\tilde{x},\tilde{x})\ , \tag{2.3}$$

where $k_{[N]}(\tilde{x}) = [k(\tilde{x}_1,\tilde{x}),\ldots,k(\tilde{x}_N,\tilde{x})]^T \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{K}_{[N]} = [k(\tilde{x}_i,\tilde{x}_j]_{i=1,j=1}^N$.

In particular, the posterior distribution of $f(x)$ is $\mathcal{N}(\mu_N(x), \sigma_N^2(x))$ and that of the corresponding observation $r(x)$ is $\mathcal{N}(\mu_N(x), \sigma_N^2(x) + \sigma^2)$.

2.5.3 Information gain

In Bayesian optimization, the informativeness of a finite sequence $\tilde{\boldsymbol{x}}_{[N]}$ is quantified by the information gain, defined as $I(\boldsymbol{r}_{[N]}; \boldsymbol{f}_{[N]}) = H(\boldsymbol{r}_{[N]}) - H(\boldsymbol{r}_{[N]}|\boldsymbol{f}_{[N]})$, where $H(\cdot)$ denotes the entropy of a random variable and $H(\cdot|\boldsymbol{f}_{[N]})$ denotes the conditional entropy of a random variable with respect to $\boldsymbol{f}_{[N]}$. The information gain gives us the decrease in entropy of $\boldsymbol{f}_{[N]}$ given the outcomes $\boldsymbol{r}_{[N]}$. We define the maximum information gain as $\gamma_N = \max_{\tilde{\boldsymbol{x}}_{[N]}} I(\boldsymbol{r}_{[N]}; \boldsymbol{f}_{[N]})$.

Note that γ_N is the maximum information gain associated with any N-tuple of elements of $\mathcal{X}$. For large (potentially infinite) spaces this quantity can be very large, depending on the chosen kernel. On the other hand, the volatile (i.e., dynamically varying) base arm availability in our setting does not require such information. We only need the maximum possible information coming from a fixed sequence of base arm (context) arrivals. Thus, such a notion becomes redundant in this case. This motivates us to relate the growth rate of the regret with the information content of the available base arms. Hence, we begin by adapting the definition of the maximum information gain so that it accommodates base arm availability, and guarantees tightness of the regret bounds. Given $t \geq 1$, let $\mathcal{Z}_t \subset 2^{\mathcal{X}_t}$ be the set of context vectors corresponding to the feasible set $\mathcal{S}_t$ of super arms[2] Let $\tilde{\boldsymbol{z}}_t := \boldsymbol{x}_{t,S}$, for some $S \in \mathcal{S}_t$, be an element of $\mathcal{Z}_t$ and let $\tilde{\boldsymbol{z}}_{[T]} = [\tilde{\boldsymbol{z}}_1^T, \ldots, \tilde{\boldsymbol{z}}_T^T]$. We define the maximum information gain associated with the sequence of context arrivals $\mathcal{X}_1, \ldots, \mathcal{X}_T$ as

$$\overline{\gamma}_T = \max_{\tilde{\boldsymbol{z}}_{[T]}: \tilde{\boldsymbol{z}}_t \in \mathcal{Z}_t, t \leq T} I(r(\tilde{\boldsymbol{z}}_{[T]}); f(\tilde{\boldsymbol{z}}_{[T]})) \,. \tag{2.4}$$

The information gain from any sequence of T selected super arms is upper bounded by $\overline{\gamma}_T$. The regret bounds of O'CLOK-UCB depend on it.

[2]Note that $\mathcal{Z}_t$ is a subset of $\mathcal{Z}$ defined in Definition 2.2.

Chapter 3

Solving CCV-MAB via adaptive discretization

3.1 The ACC-UCB algorithm

Our first method is called *Adaptive Contextual Combinatorial Upper Confidence Bound* (ACC-UCB) and is motivated by several tree based methods that have been used for function optimization under continuity assumptions [35, 37, 45] (pseudocode given in Algorithm 2). We assume that the context space $(\mathcal{X}, d)$ is well behaved. Thus, by Definition 2.1, there exists a sequence $(\mathcal{X}_h)_{h\geq 0}$, each element of which contains N^h nodes whose associated cells form a tree of partitions of $\mathcal{X}$. The procedure is described as follows: at round t, we observe arrived base arms and contexts. We maintain an active set of leaf nodes and denote it by $\mathcal{L}_t$. For the arrived base arms, we identify the set of available active leaf nodes, whose regions contain the available contexts and denote it by $\mathcal{N}_t$. By $par(\chi_{h,i})$ we denote the parent of node $\chi_{h,i}$. Fig. 3.1 illustrates the nodes and cells of a 1D adaptive discretization. For each active leaf node we maintain an index which is an upper confidence bound on the maximum expected outcome of the base arms that have contexts in the region associated with the node. The AD-index

(the f estimate using adaptive discretization) is defined as

$$I_t^{ad}(\chi_{h,i}) \coloneqq \vartheta_t(\chi_{h,i}) + v_1\rho^h$$

where the term $\vartheta_t(\chi_{h,i})$ is a high probability upper bound on $f(\chi_{h,i})$ defined as

$$\vartheta_t(\chi_{h,i}) \coloneqq \min\{\hat{\mu}_{t-1}(\chi_{h,i}) + c_{t-1}(\chi_{h,i}), \hat{\mu}_{t-1}(par(\chi_{h,i})) + c_{t-1}(par(\chi_{h,i})) + v_1\rho^{(h-1)}\}$$

and $c_t(\chi_{h,i}) \coloneqq \sqrt{2\log(\sqrt{K\sqrt{2N}}T)/C_t(\chi_{h,i})}$ is the confidence radius, tailored to give high probability upper bounds on f[1]. Here, $C_t(\chi_{h,i})$ is the number of times a base arm associated with a context from the cell $X_{h,i}$ was selected by the MCSP, formally defined as

$$C_t(\chi_{h,i}) \coloneqq \sum_{t'=1}^{t}\sum_{k=1}^{|S_t|}\mathbb{I}\{(H_{t',k}, I_{t',k}) = (h,i)\}$$

where we denote by $(H_{t',k}, I_{t',k})$ level and AD-index of the active leaf node associated with the cell containing the context of the kth selected base arm at time t', and let $\phi_{t,k} \coloneqq \chi_{H_{t,k},I_{t,k}}$. One thing to be noted here is that the term $\vartheta_t(\chi_{h,i})$ takes the minimum of two upper bounds on the value of the true mean at $\chi_{h,i}$. First, note that $f(\chi_{h,i})$ can be upper bounded by $\hat{\mu}_{t-1}(\chi_{h,i}) + c_{t-1}(\chi_{h,i})$, using Hoeffding bounds. Similarly, the second term also upper bounds $f(\chi_{h,i})$, since $\chi_{h,i}$ is also inside the cell centered at $par(\chi_{h,i})$, and therefore, due to Lipschitz continuity of f, its value at $\chi_{h,i}$ will not exceed the previous term plus the diameter of the cell. Taking the minimum thus guarantees the best of the upper bounds.

We define the total reward accumulated by the algorithm until round t from selecting base arms with contexts associated with the node $\chi_{h,i}$ as follows:

$$\upsilon_t(\chi_{h,i}) \coloneqq \sum_{t'=1}^{t}\sum_{k=1}^{|S_t|} r(x_{t',s_{t',k}})\mathbb{I}\{(H_{t',k}, I_{t',k}) = (h,i)\}$$

[1]Note that here we assume the learner knows K beforehand. In practice, this is usually the case. The crowdsourcing experiments that we perform allow for such an assumption. The reason we use K instead of M is that in reality M can be in the orders of millions, while K may be less than 50.

where we denote by $s_{t,k}$ the kth base arm selected by the algorithm at round t. Consequently, we define the empirical mean used in the AD-index as

$$\hat{\mu}_t(\chi_{h,i}) \coloneqq \begin{cases} v_t(\chi_{h,i})/C_t(\chi_{h,i}) & \text{for } C_t(\chi_{h,i}) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that $C_t(\chi_{h,i})$ can be larger than t, since at a certain round the the algorithm may choose base arms with contexts belonging to the same cell. However, it always holds that $C_t(\chi_{h,i}) \leq Kt$. The constants v_1 and ρ are parameters as described in Definition 2.1 that are given as input to the algorithm.

Next, we define the AD-index of a base arm $m \in \mathcal{M}_t$ at round t. For this, let $(\tilde{H}_{t,m}, \tilde{I}_{t,m})$ represent level and AD-index of the active leaf node associated with the cell containing the context $x_{t,m}$, and let $\tilde{\phi}_{t,m} \coloneqq \chi_{\tilde{H}_{t,m},\tilde{I}_{t,m}}$. We define the AD-index of a base arm $m \in \mathcal{M}_t$ as

$$I_t^{ad}(x_{t,m}) \coloneqq I_t^{ad}(\tilde{\phi}_{t,m}) + N(v_1/v_2)v_1\rho^{\tilde{H}_{t,m}}$$

where the second term guarantees (with high probability) that $I_t^{ad}(x_{t,m})$ upper bounds $f(x_{t,m})$.

Remark 2. *Since at any round t, a cell associated with an active leaf node $\chi_{h,i}$ may contain several contexts of the available base arms, we have $I_t^{ad}(x_{t,m}) = I_t^{ad}(x_{t,n})$ when m and n are two available base arms, both of which have contexts that live in the cell $X_{h,i}$. As a consequence, indices of all base arm contexts inside one cell are equal.*

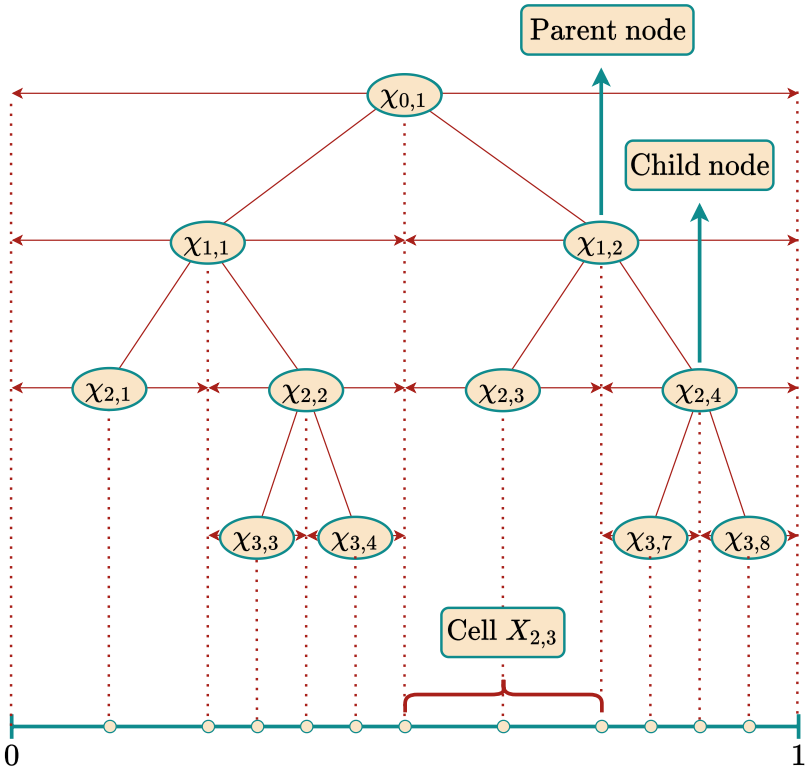

Figure 3.1: Illustration of adaptive discretization in the case when $\mathcal{X} = [0, 1]$.

After the indices of the available base arms, i.e., $\{I_t^{ad}(x_{t,m})\}_{m \in \mathcal{M}_t}$ are computed, they are given as input $\boldsymbol{\theta}_t$ to the approximation oracle in round t to obtain the super arm $S_t \subset \mathcal{M}_t$ that will be played in round t.[2] At this point, we identify the active leaf nodes that are "selected", denote their collection by $\mathcal{P}_t$, and update their statistics (after these are played and their outcomes are observed) according to the following rules. For each $\chi_{h,i} \in \mathcal{P}_t$:

$$\hat{\mu}_t(\chi_{h,i}) = \frac{C_{t-1}(\chi_{h,i})\hat{\mu}_{t-1}(\chi_{h,i}) + rew_t(\chi_{h,i})}{C_{t-1}(\chi_{h,i}) + num_t(\chi_{h,i})} \tag{3.1}$$

$$C_t(\chi_{h,i}) = C_{t-1}(\chi_{h,i}) + num_t(\chi_{h,i}) \ . \tag{3.2}$$

where

$$rew_t(\chi_{h,i}) := \sum_{k=1}^{|S_t|} r(x_{t',s_{t',k}}) \mathbb{I}\{(H_{t',k}, I_{t',k}) = (h, i)\}$$

[2]The base arms are randomly chosen in the first round.

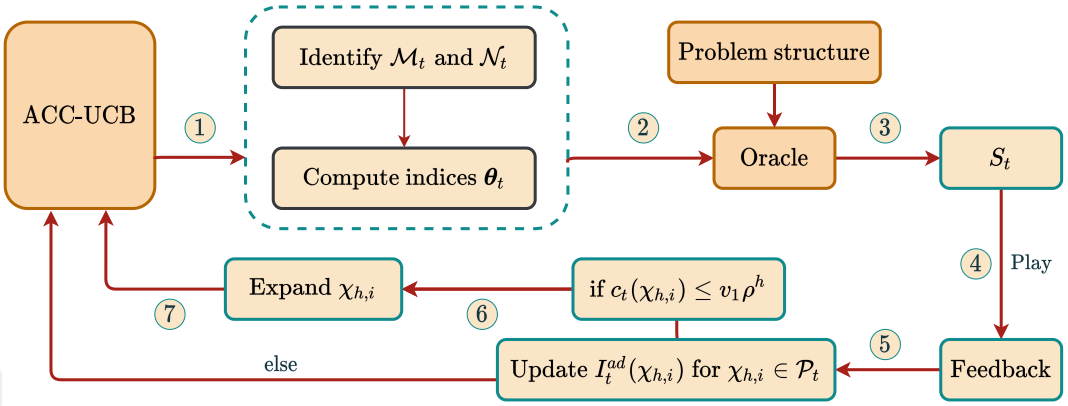

Figure 3.2: Illustration of the ACC-UCB algorithm. The learner identifies $\mathcal{M}_t$ and $\mathcal{N}_t$ and thereby computes the indices $\boldsymbol{\theta}_t$ which he then feeds to the oracle. The oracle then returns the α-optimal super arm S_t, which the learner plays and from which he consequently observes semi-bandit feedback. Using this feedback, the indices of the nodes in $\mathcal{P}_t$ corresponding to the played super arm are updated. Finally, the learner decides whether any nodes in $\mathcal{P}_t$ need to be expanded and then proceeds to the next round.

$$num_t(\chi_{h,i}) := \sum_{k=1}^{|S_t|} \mathbb{I}\{(H_{t',k}, I_{t',k}) = (h,i)\} \; .$$

Statistics of the other active leaf nodes do not change. Subsequently, for each node $\chi_{h,i} \in \mathcal{P}_t$, we decide whether or not to expand it into N children nodes, according to the following condition:

- *Refine*. If $c_t(\chi_{h,i}) \leq v_1 \rho^h$, then the node $\chi_{h,i}$ is expanded into N children nodes $\{\chi_{h+1,j} \; : \; N(i-1)+1 \leq j \leq Ni\}$ which are added to the set of active leaves, whereas $\chi_{h,i}$ is removed from it.

If the above condition is not satisfied, we continue. Basically, we refine the partitions when we are confident enough about the f values inside that cell. Fig. 3.2 visualizes the steps of our algorithm.

Algorithm 1 ACC-UCB

Input: $\mathcal{X}$, $(\mathcal{X}_h)_{h\geq 0}$, v_1, v_2, ρ, K, N, T.
  Initialize: $C_0(\chi_{h,i}) = 0$, $\hat{\mu}_0(\chi_{h,i}) = 0$, $\forall \chi_{h,i} \in \mathcal{X}$; $\mathcal{X}_0 = \mathcal{X}$, $\mathcal{L}_1 = \{\chi_{0,1}\}$.
  for $t = 1, 2, \ldots, T$ **do**:
    Observe base arms in $\mathcal{M}_t$ and their contexts $\mathcal{X}_t$.
    Identify available active leaf nodes $\mathcal{N}_t \subseteq \mathcal{L}_t$.
    Compute indices $I_t^{ad}(\chi_{h,i})$, of nodes $\chi_{h,i} \in \mathcal{N}_t$.
    Compute indices $I_t^{ad}(x_{t,m})$ of contexts $x_{t,m} \in \mathcal{X}_t$.
    $S_t \leftarrow$ Oracle$((I_t^{ad}(x_{t,m}))_{m\in\mathcal{M}_t}, \mathcal{S}_t, M_t)$
    Observe outcomes of base arms in S_t and collect the reward.
    Identify the set of selected nodes $\mathcal{P}_t$.
    for $\chi_{h,i} \in \mathcal{P}_t$ **do**:
      Update $\hat{\mu}_t(\chi_{h,i})$ as in (3.1) and $C_t(\chi_{h,i})$ as in (3.2).
      if $c_t(\chi_{h,i}) \leq v_1\rho^h$ **then**: ▷ Refine
        $\mathcal{L}_{t+1} \leftarrow \mathcal{L}_t \cup \{\chi_{h+1,j} : \ N(i-1)+1 \leq j \leq Ni\} \setminus \{\chi_{h,i}\}$.
      end if
    end for
  end for

3.2 Regret analysis

In this section we state and prove the regret bounds of ACC-UCB. Furthermore, we lay down another approach to tighter bounds, by making use of benign context arrivals. In order for our bounds to hold, we need to make a further assumption on the expected base arm outcomes. We assume that base arms with similar contexts have similar expected outcomes. This is captured by imposing the following smoothness condition on f, which is a standard assumption in the contextual MAB setting [13].

Assumption 3. *(Lipschitz continuity for the expected outcome in contexts). For any given pair of contexts $x, x' \in \mathcal{X}$, we have $|f(x) - f(x')| \leq d(x, x')$.*

Note that this assumption is substituted with the smoothness of f induced from the GP structure in the O'CLOK-UCB's case.

3.2.1 General regret bounds

First, we show that the AD-index of a given node $\chi_{h,i}$ is an upper bound of its true mean with overwhelming probability.

Lemma 3.1. *Given the event*

$$\mathcal{F} = \left\{\forall t \leq T,\, \forall \chi_{h,i} \in \mathcal{L}_t :\, |\hat{\mu}_t(\chi_{h,i}) - f(\chi_{h,i})| \leq c_t(\chi_{h,i}) + v_1\rho^h\right\}$$

under the assumptions made so far, we have $\mathbb{P}\{\mathcal{F}\} \geq 1 - 1/T$.

Proof. Let $M(t) = \sum_{i=1}^{t} |S_i|$. We represent the context of the kth selected base arm in round t by $X_{M(t-1)+k}$. We represent the observed outcome related to X_t by Y_t and the active leaf node associated with the cell that contains X_t by (H_t, I_t). Let $M_j = \min\{s \in \mathbb{N} :\ \sum_{i=1}^{s} |S_i| \geq j\}$ represent the round in which the total number of selected base arms reach j. Assume that all random variables $X_1, Y_1, X_2, Y_2, \ldots$ are defined over the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and

$$\mathcal{F}_t = \sigma(X_1, X_2, \ldots, X_t, \ldots, X_{t+|S_{M_t}|}, Y_1, \ldots, Y_t)$$

$\mathbb{F} = (\mathcal{F}_t)_{t=0}^{\infty}$ is an $\mathcal{F}$-adapted filtration. We have that $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = f(X_t)$.

Fix a node $\chi_{h,i}$. Let θ_t be the indicator variable that represents the event that X_t is in the cell associated with node $\chi_{h,i}$ after $\chi_{h,i}$ was created (this also counts the times when X_t is in $X_{h,i}$ after $\chi_{h,i}$ is expanded into children nodes). When $\theta_t = 1$, we say that $\chi_{h,i}$ is *observed*. Let $Z_0 = 0$ and

$$Z_t = \sum_{j=1}^{t} (f(X_j) - Y_j)\theta_j\ .$$

Note that Z_t is $\mathcal{F}_t$-measurable. Moreover, X_t and θ_t are all $\mathcal{F}_{t-1}$-measurable while Y_t is is not. Thus, we have

$$\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = \mathbb{E}\left[(f(X_t) - Y_t)\theta_t\,|\mathcal{F}_{t-1}\right] + \mathbb{E}\left[\sum_{j=1}^{t-1} (f(X_j) - Y_j)\theta_j\,|\mathcal{F}_{t-1}\right]$$

$$
\begin{aligned}
&= (f(X_t) - \mathbb{E}[Y_t|\mathcal{F}_{t-1}])\theta_t + \sum_{j=1}^{t-1}(f(X_j) - Y_j)\theta_j \\
&= (f(X_t) - f(X_t))\theta_t + Z_{t-1} \\
&= Z_{t-1} \ .
\end{aligned}
$$

This shows that $(Z_t)_{t=1}^{\infty}$ is an $\mathbb{F}$-adapted martingale. Let

$$
\tilde{C}_t(\chi_{h,i}) = \sum_{j=1}^{t} \theta_j \ .
$$

We define a sequence of random stopping times when the base arm outcomes associated with nodes $\chi_{h,i}$ were *observed*: $T_j = \min\{l : \tilde{C}_l(\chi_{h,i}) = j\}$. Note that $\{T_j = t\}$ is $\mathcal{F}_{t-1}$-measurable for all t and (T_j) is an increasing sequence of stopping times, i.e., $1 \leq T_1 < T_2 < \ldots$, hence it holds that $T_j \geq j$. Our next aim is to apply Doob's optional skipping theorem [46, Theorem 2.3]. As its application requires $T_1 < T_2 < \ldots < \infty$, we assume that after T there are infinitely many rounds in which all contexts in everyone of them arrive in $\chi_{h,i}$ (this ensures that base arms with contexts in $\chi_{h,i}$ are observed infinitely often). Note that this technical assumption has no effect on our model as it happens after our horizon of interest. Let $\tilde{Z}_j = Z_{T_j}$. Then, by Doob's optional skipping theorem $(\tilde{Z}_j)$ is a martingale w.r.t. $(\mathcal{F}_{T_j})$.

In this proof only, with an abuse of notation, let (H_t, I_t) represent the level and the AD-index of the leaf node that contains X_t. We denote by $\tilde{X}_j = X_{T_j}$ the jth base arm selected in the region corresponding to $\chi_{h,i}$ and $\tilde{Y}_j = Y_{T_j}$. Let $n = C_t(\chi_{h,i})$. Note that when $\chi_{h,i} \in \mathcal{L}_t$, we have $\tilde{C}_{M(t)}(\chi_{h,i}) = C_t(\chi_{h,i})$ and $\theta_j = \mathbb{I}((H_j, I_j) = (h, i))$. We have

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{j=1}^{t}\sum_{k=1}^{|S_t|}(f(x_{j,s_{j,k}}) - p(x_{j,s_{j,k}}))\mathbb{I}((H_{j,k}, I_{j,k}) = (h,i))\right| \geq \sqrt{2n\log(\sqrt{K\sqrt{2N}T})},\right. \\
&\quad \left.\chi_{h,i} \in \mathcal{L}_t\right\} \qquad (3.3) \\
&= \sum_{l=1}^{M(t)} \mathbb{P}\left\{\left|\sum_{j=1}^{t}\sum_{k=1}^{|S_t|}(f(x_{j,s_{j,k}}) - p(x_{j,s_{j,k}}))\mathbb{I}((H_{j,k}, I_{j,k}) = (h,i))\right|\right.
\end{aligned}
$$

$$\geq \sqrt{2n\log(\sqrt{K\sqrt{2N}T})}, \chi_{h,i} \in \mathcal{L}_t, n = l \Bigg\} \tag{3.4}$$

$$= \sum_{l=1}^{M(t)} \mathbb{P}\left\{ \left| \sum_{j=1}^{M(t)} (f(X_j) - Y_j)\mathbb{I}(H_j, I_j) = (h,i) \right| \geq \sqrt{2n\log(\sqrt{K\sqrt{2N}T})}, \right.$$

$$\left. \chi_{h,i} \in \mathcal{L}_t, \tilde{C}_{M(t)}(\chi_{h,i}) = l \right\} \tag{3.5}$$

$$= \sum_{l=1}^{M(t)} \mathbb{P}\left\{ \left| \sum_{j=1}^{l} (f(\tilde{X}_j) - \tilde{Y}_j) \right| \geq \sqrt{2l\log(\sqrt{K\sqrt{2N}T})}, \chi_{h,i} \in \mathcal{L}_T, \tilde{C}_{M(t)}(\chi_{h,i}) = l \right\} \tag{3.6}$$

$$= \sum_{l=1}^{M(t)} \mathbb{P}\left\{ |\tilde{Z}_l| \geq \sqrt{2l\log(\sqrt{K\sqrt{2N}T})}, \chi_{h,i} \in \mathcal{L}_t, \tilde{C}_{M(t)}(\chi_{h,i}) = l \right\} \tag{3.7}$$

$$\leq \sum_{l=1}^{M(t)} \mathbb{P}\left\{ \left|\tilde{Z}_l\right| \geq \sqrt{2l\log(\sqrt{K\sqrt{2N}T})} \right\} \tag{3.8}$$

$$\leq \sum_{l=1}^{M(t)} 2\text{exp}\left(- \frac{2\left(\sqrt{2l\log(\sqrt{K\sqrt{2N}T})}\right)^2}{l} \right) \tag{3.9}$$

$$= 2KT \cdot \frac{1}{2} N^{-1}K^{-2}T^{-4} = N^{-1}K^{-1}T^{-3} \tag{3.10}$$

where (3.4) follows from the law of total probability and for (3.9) we use the Azuma-Hoeffding inequality for martingale differences noting that the outcomes lie in the unit interval. Therefore, with probability at least $1 - N^{-1}K^{-1}T^{-3}$ we have

$$\sum_{j=1}^{t} \sum_{k=1}^{M(t)} (p(x_{j,s_{j,k}}) - f(x_{j,s_{j,k}}))\mathbb{I}((H_{j,k}, I_{j,k}) = (h,i)) \leq \sqrt{2n\log(\sqrt{K\sqrt{2N}T})}\,.$$

Next, by Definition 2.1 and Assumption 3 when $\mathbb{I}((H_{j,k}, I_{j,k}) = (h,i)) = 1$, we have that $f(x_{j,s_{j,k}}) - f(\chi_{h,i}) \leq v_1\rho^h$ since the context lives in $X_{h,i}$. Consequently, by

using the triangle inequality, it follows that

$$\begin{aligned}
&\Big|\sum_{j=1}^{t}\sum_{k=1}^{M(t)}(p(x_{j,s_{j,k}})-f(\chi_{h,i}))\mathbb{I}((H_{j,k},I_{j,k})=(h,i))\Big| \\
&\leq \Big|\sum_{j=1}^{t}\sum_{k=1}^{M(t)}(p(x_{j,s_{j,k}})-f(x_{j,s_{j,k}}))\mathbb{I}((H_{j,k},I_{j,k})=(h,i))\Big| \\
&\qquad\qquad + \Big|\sum_{j=1}^{t}\sum_{k=1}^{M(t)}(f(x_{j,s_{j,k}})-f(\chi_{h,i}))\mathbb{I}((H_{j,k},I_{j,k})=(h,i))\Big| \\
&\leq \sqrt{2n\log(\sqrt{K\sqrt{2N}T})}+nv_1\rho^h\ .
\end{aligned}$$

Dividing both sides by n, we obtain

$$|\hat{\mu}_t(\chi_{h,i})-f(\chi_{h,i})|\leq c_t(\chi_{h,i})+v_1\rho^h\ .$$

Note that until round T, the maximal number of possible new node activations is NKT and therefore for the above event to hold simultaneously for all nodes at all times, we take the union bound over all nodes and over all times. From this, we obtain

$$\mathbb{P}\big\{\forall t\leq T,\,\forall\chi_{h,i}\in\mathcal{L}_t:\,|\hat{\mu}_t(\chi_{h,i})-f(\chi_{h,i})|\leq c_t(\chi_{h,i})+v_1\rho^h\big\}\geq 1-1/T\ .$$

■

The next result gives high probability bounds of the difference between the AD-index and the mean of a given node. Furthermore, we give an upper bound on the number of times a node can be selected before expansion.

Lemma 3.2. *Consider that event $\mathcal{F}$ happens. Then, if at round t, the node $\chi_{h,i}\in\mathcal{P}_t$ is not expanded by the algorithm, we have*

$$I_t^{ad}(\chi_{h,i})-f(\chi_{h,i})\leq(5Nv_1/v_2+1)v_1\rho^h.$$

Moreover, a node $\chi_{h,i}$ may be selected by the algorithm no more than q_h times before

it is expanded, where

$$q_h \leq \left\lceil \frac{2\log(Tv_3)}{(v_1\rho^h)^2} \right\rceil \leq \frac{2\log(Tv_3)}{(v_1\rho^h)^2} \ ,$$

with $v_3 := \sqrt{\sqrt{2N}Ke^{v_1^2}}$.

Proof. The following inequalities hold.

$$\begin{aligned}
I_t^{ad}(\chi_{h,i}) &= \vartheta_t(\chi_{h,i}) + v_1\rho^h && (3.11)\\
&\leq \hat{\mu}_{t-1}(par(\chi_{h,i})) + c_{t-1}(par(\chi_{h,i})) + v_1\rho^{(h-1)} + v_1\rho^h && (3.12)\\
&\leq f(par(\chi_{h,i})) + 2c_{t-1}(par(\chi_{h,i})) + 2v_1\rho^{(h-1)} + v_1\rho^h && (3.13)\\
&\leq (f(par(\chi_{h,i})) - v_1\rho^{(h-1)}) + 5v_1\rho^{(h-1)} + v_1\rho^h && (3.14)\\
&\leq f(\chi_{h,i}) + 5v_1\rho^{(h-1)} + v_1\rho^h && (3.15)\\
&\leq f(\chi_{h,i}) + (5Nv_1/v_2 + 1)v_1\rho^h \ , && (3.16)
\end{aligned}$$

where the inequality in (3.12) follows from the definition of $\vartheta_t(\chi_{h,i})$, while (3.13) follows from Lemma 3.1. The inequality in (3.14) follows from the fact that $par(\chi_{h,i})$ must have been expanded, thus we must have had $c_{t-1}(par(\chi_{h,i})) \leq v_1\rho^{(h-1)}$. For the inequality in (3.15) we use the Lipschitz continuity for the means of nodes and the fact that $\chi_{h,i}$ and $par(\chi_{h,i})$ lie in the cell associated with $par(\chi_{h,i})$. Lastly, the inequality in (3.16) follows from the triangle inequality. By Definition 2.1, the cell $X_{h-1,i}$ is expanded into N disjoint cells of diameter less than $v_1\rho^h$, for any $1 \leq i \leq N^{h-1}$, therefore $v_2\rho^{h-1} \leq Nv_1\rho^h$.

Finally, solving $c_T(\chi_{h,i}) = v_1\rho^h$ gives $C_T(\chi_{h,i}) = 2\log(\sqrt{K\sqrt{2N}}T)/(v_1\rho)^2$. Thus, we must have

$$\begin{aligned}
C_T(\chi_{h,i}) &\leq \left\lceil \frac{2\log(\sqrt{K\sqrt{2N}}T)}{(v_1\rho)^2} \right\rceil \leq \frac{2\log(\sqrt{K\sqrt{2N}}T)}{(v_1\rho)^2} + 1\\
&\leq \frac{2\log T + 2(\log\sqrt{K\sqrt{2N}}) + v_1^2}{(v_1\rho)^2}\\
&= \frac{2\log T + 2\log(\sqrt{\sqrt{2N}Ke^{v_1^2}})}{(v_1\rho)^2}
\end{aligned}$$

$$= \frac{2\log(Tv_3)}{(v_1\rho)^2} ,$$

where $v_3 = \sqrt{\sqrt{2N}Ke^{v_1^2}}$. ■

Next, we give a high probability upper bound on the true mean of any given base arm.

Lemma 3.3. *Consider that event $\mathcal{F}$ happens. Then, we have $\forall t \leq T$ and $m \in \mathcal{M}_t$, $I_t^{ad}(x_{t,m}) \geq f(x_{t,m})$.*

Proof. First, note that we have $N(v_1/v_2) \geq 1$ (since $N \geq 2$ and $v_2 \leq 1 \leq v_1$). Now, let $\tilde{\phi}_{t,m} = \chi_{h,i}$, for some $h, i \in \mathbb{N} \cup \{0\}$. We have

$$I_t^{ad}(x_{t,m}) = I_t^{ad}(\chi_{h,i}) + N(v_1/v_2)v_1\rho^h .$$

We now have on event $\mathcal{F}$

$$\begin{aligned} I_t^{ad}(x_{t,m}) &= I_t^{ad}(\chi_{h,i}) + N(v_1/v_2)v_1\rho^h \\ &= \vartheta_t(\chi_{h,i}) + v_1\rho^h + N(v_1/v_2)v_1\rho^h \\ &= \min\{\hat{\mu}_{t-1}(\chi_{h,i}) + c_{t-1}(\chi_{h,i}), \hat{\mu}_{t-1}(par(\chi_{h,i})) + c_{t-1}(par(\chi_{h,i})) + v_1\rho^{(h-1)}\} \\ &\quad + v_1\rho^h + N(v_1/v_2)v_1\rho^h . \end{aligned}$$

We have two cases. If the minimum is $\hat{\mu}_{t-1}(\chi_{h,i}) + c_{t-1}(\chi_{h,i})$, then we have

$$\begin{aligned} I_t^{ad}(x_{t,m}) &= \hat{\mu}_{t-1}(\chi_{h,i}) + c_{t-1}(\chi_{h,i}) + v_1\rho^h + N(v_1/v_2)v_1\rho^h \\ &\geq \hat{\mu}_{t-1}(\chi_{h,i}) + c_{t-1}(\chi_{h,i}) + 2v_1\rho^h \\ &\geq f(\chi_{h,i}) + v_1\rho^h \\ &\geq f(x_{t,m}) , \end{aligned}$$

since event $\mathcal{F}$ holds, $N(v_1/v_2) \geq 1$, and also by Lipschitz continuity and Definition 2.1.

If the minimum is $\hat{\mu}_{t-1}(par(\chi_{h,i})) + c_{t-1}(par(\chi_{h,i})) + v_1\rho^{(h-1)}$, then we have

$$\begin{aligned} I_t^{ad}(x_{t,m}) &= \hat{\mu}_{t-1}(par(\chi_{h,i})) + c_{t-1}(par(\chi_{h,i})) + v_1\rho^{(h-1)} + N(v_1/v_2)v_1\rho^h + v_1\rho^h \\ &\geq f(par(\chi_{h,i})) + N(v_1/v_2)v_1\rho^h + v_1\rho^h \\ &\geq f(par(\chi_{h,i})) + v_1\rho^{h-1} + v_1\rho^h \\ &\geq f(\chi_{h,i}) + v_1\rho^h \\ &\geq f(x_{t,m}) \ , \end{aligned}$$

since $N(v_1/v_2)v_1\rho^h \geq v_1\rho^{h-1}$ and $|f(par(\chi_{h,i})) - f(\chi_{h,i})| \leq v_1\rho^{h-1}$. ■

The following lemma upper bounds the suboptimality gap of any suboptimal super arm with high probability.

Lemma 3.4. *If the reward function u satisfies the Lipschitz continuity and the monotonicity condition and if event $\mathcal{F}$ holds, then in any round t, the regret incurred by any suboptimal super arm S_t is upper bounded by*

$$\sum_{m \in S_t, \chi_{h,i} \in \mathcal{L}_t : \tilde{\phi}_{t,m} = \chi_{h,i}} B(6Nv_1/v_2 + 2)v_1\rho^h$$

or by $BK(6Nv_1/v_2 + 2)v_1\rho^{h(t)}$, where

$$h(t) = \min\{h : \chi_{h,i} \in \mathcal{L}_t, \tilde{\phi}_{t,m} = \chi_{h,i},\ m \in S_t\} \ .$$

Proof. Let $I^{ad}(\boldsymbol{x}_{t,S^t}) = [I_t^{ad}(x_{t,m})]_{m \in S_t}$ be the vector of indices of base arms in S_t. Then we have:

$$u(I^{ad}(\boldsymbol{x}_{t,S_t})) \geq \alpha \cdot u(I^{ad}(\boldsymbol{x}_{t,G_t})) \geq \alpha \cdot u(I^{ad}(\boldsymbol{x}_{t,S_t^*})) \geq \alpha \cdot u(f(\boldsymbol{x}_{t,S_t^*})) \ ,$$

where $G_t := \text{argmax}_{S \in \mathcal{S}_t} u(I^{ad}(\boldsymbol{x}_{t,S}))$. The first inequality holds because we choose S_t by the α-approximation oracle; the second inequality follows from the definition of G_t; the third inequality follows from Lemma 3.3 and monotonicity, since $g(x_{t,m}) \geq f(x_{t,m})$, for all $m \in S_t^*$. Letting $\Delta(S_t)$ be the approximate optimality gap of S_t, we now have

$$
\begin{aligned}
\Delta(S_t) &= \alpha \cdot \text{opt}(\boldsymbol{f}_t) - u(f(\boldsymbol{x}_{t,S_t})) \\
&\leq u(g(\boldsymbol{x}_{t,S_t})) - u(f(\boldsymbol{x}_{t,S_t})) && (3.17)\\
&\leq B \cdot \sum_{m \in S_t} |I_t^{ad}(x_{t,m}) - f(x_{t,m})| && (3.18)\\
&\leq B \cdot \sum_{\substack{h:\, \chi_{h,i} \in \mathcal{L}_t \\ \tilde{\phi}_{t,m} = \chi_{h,i},\, m \in S_t}} |I_t^{ad}(\chi_{h,i}) + N(v_1/v_2)v_1\rho^h - f(\chi_{h,i})| + |f(\chi_{h,i}) - f(x_{t,m})| && (3.19)\\
&\leq B \cdot \sum_{\substack{h:\, \chi_{h,i} \in \mathcal{L}_t \\ \tilde{\phi}_{t,m} = \chi_{h,i},\, m \in S_t}} |I_t^{ad}(\chi_{h,i}) - f(\chi_{h,i})| + v_1\rho^h + N(v_1/v_2)v_1\rho^h && (3.20)\\
&\leq B \cdot \sum_{m \in S_t} (6Nv_1/v_2 + 2)v_1\rho^{\tilde{h}_{t,m}} && (3.21)\\
&\leq BK(6Nv_1/v_2 + 2)v_1\rho^{h(t)}\,, && (3.22)
\end{aligned}
$$

where (3.17) follows from the above argument; (3.18) follows from the Lipschitz continuity of the expected reward function; in (3.19) we have added and subtracted the true mean of the context $\chi_{h,i}$ (which represents a node), and than used the triangle inequality; in (3.20) we have used the Lipschitz continuity for the expected outcomes and Definition 2.1, in the sense that any two contexts in the same cell are at most a diameter away from each other; (3.21) follows from Lemma 3.2 and (3.22) follows from the way we have defined $h(t)$. ■

Now we are ready to prove our main result.

Theorem 3.1. *Fix $T > 0$. Given the parameters of the problem $0 < \alpha \leq 1$, $N \in \mathbb{N}$, $K \in \mathbb{N}$, $B > 0$ and $0 < v_2 \leq 1 \leq v_1$, define $\bar{D} = D^u(g, \alpha u^*_{\min})$ and $g(r) = cr$, where $c = BK(6Nv_1/v_2 + 2)v_1/v_2$. Then, for any $D_1 > \bar{D}$, there exists $Q = Q(\mathcal{X}, u, g, \alpha u^*_{\min}, c) > 0$ (independent of T), for which the α-regret incurred by ACC-UCB is upper bounded with probability at least $1 - 1/T$ as follows:*

$$
R_\alpha(T) \leq (C_1 + C_2) \cdot T^{1 - \frac{1}{D_1+2}} \cdot (\log(Tv_3))^{\frac{1}{D_1+2}}
$$

where $C_1 := 2QBK(6Nv_1/v_2+2)\frac{v_2^{-D_1}}{v_1(\rho^{-1}-1)}$ *and* $C_2 := KB(6Nv_1/v_2+2)v_1$.

Proof. By Lemma 3.4, at any given round t, the algorithm only selects contexts from $(BK(6Nv_1/v_2+2)v_1\rho^{h(t)}, u, \alpha u^*_{\min})$-optimal sets. Note that for all h, we have $|\mathcal{X}_h| = N^h$ by Definition 2.1. Furthermore, we select only contexts from sets of the form $\mathcal{X}^{\alpha u^*_{\min}}_{BK6Nv_1/v_2+2)v_1\rho^{h(t)}}$, for all $t \geq 0$. From Lemma 3.4, we have that

$$R_\alpha(T) \leq \sum_{t\leq T}\sum_{k=1}^{K} B(6Nv_1/v_2+2)v_1\rho^{H_{t,k}} \leq \sum_{t\leq T} BK(6Nv_1/v_2+2)v_1\rho^{h(t)} \ .$$

Note that in each round, we consider the contribution that comes from the highest leaf node in the tree (i.e., the leaf node with the smallest h) associated with the selected super arm. In order to obtain sublinear regret, we express the summation in terms of the levels of the tree. In order to do that, we observe that we may have $h(t) = h(t')$ for $t \neq t'$. That is for several reasons. First, we have up to $N^{h(t)}$ nodes in a certain level, but the learner does not select contexts associated to all those nodes, only the $BK(6Nv_1/v_2+2)v_1\rho^{h(t)}$-optimal ones (more generally, the $BK(6Nv_1/v_2+2)v_1\rho^{h(t)}$-optimal ones). But since we want to define a notion of regret that captures volatility of the base arms and structure of the expected reward function, we may think that the learner selects contexts from the set $\mathcal{X}^{\alpha u^*_{\min}}_{BK(6Nv_1/v_2+2)v_1\rho^{h(t)}}$, as given in Definition 2.2. Second, the learner may select a node several times, up to q_h times by Lemma 3.2, until it is expanded. Using this approach, we will consider splitting the regret into two sums.

We fix a positive integer H. We will first bound the regret coming from rounds t such that $h(t) < H$, which we will denote by $R^1_\alpha(T)$, and the rounds t such that $h(t) \geq H$ by $R^2_\alpha(T)$. We have

$$R_\alpha(T) \leq R^1_\alpha(T) + R^2_\alpha(T) \ .$$

We bound $R^1_\alpha(T)$ first. Letting $\bar{D} = D^u(g, \alpha u^*_{\min})$ be the $(g, u, \alpha u^*_{\min})$-optimality dimension, where $g(r) = BK(6Nv_1/v_2+2)(v_1/v_2)r$, then for any $D_1 > \bar{D}$, there

exists $Q > 0$ (by Lemma 2.1), for which we have

$$
\begin{aligned}
R_\alpha^1(T) &:= \sum_{\substack{t\le T:\\ h(t)<H}} BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^{h(t)} \\
&\le \sum_{\substack{t\le T:\\ h(t)<H}} BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^{h(t)} \\
&= \sum_{h=0}^{H-1}\sum_{t\le T} \mathbb{I}(h(t)=h)BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^{h} && (3.23)\\
&\le \sum_{h=0}^{H-1} |\mathcal{X}_h \cap \mathcal{X}^{\alpha u^*_{\min}}_{BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^h}| \cdot BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^{h}\cdot q_h && (3.24)\\
&\le \sum_{h=0}^{H-1} M(\mathcal{X}^{\alpha u^*_{\min}}_{BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^h}, d, v_2\rho^h)\cdot BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^{h}\cdot q_h && (3.25)\\
&\le \sum_{h=0}^{H-1} Q\cdot (v_2\rho^h)^{-D_1} BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^{h}\cdot \frac{2\log(Tv_3)}{(v_1\rho^h)^2} && (3.26)\\
&= \sum_{h=0}^{H-1} QBK(6Nv_1/v_2+2)\frac{(v_2)^{-D_1}}{v_1}\cdot \frac{2\log(Tv_3)}{\rho^{h(D_1+1)}} && (3.27)\\
&= 2QBK(6Nv_1/v_2+2)\frac{{v_2}^{-D_1}}{v_1}\log(Tv_3)\cdot \sum_{h=0}^{H-1}\frac{1}{\rho^{h(D_1+1)}} && (3.28)\\
&\le 2QBK(6Nv_1/v_2+2)\frac{{v_2}^{-D_1}}{v_1}\frac{1-(\rho^{-1})^{H(D_1+1)}}{(1-\rho^{-1})}\cdot\log(Tv_3) && (3.29)\\
&\le 2QBK(6Nv_1/v_2+2)\frac{{v_2}^{-D_1}}{v_1(\rho^{-1}-1)}\rho^{-H(D_1+1)}\cdot\log(Tv_3)\,. && (3.30)
\end{aligned}
$$

For (3.24) we argue as follows. We have expressed the summation in terms of levels of the tree. For a certain level h, in T rounds, the learner may have selected up to $|\mathcal{X}_h \cap \mathcal{X}^{\alpha u^*_{\min}}_{BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^h}|$ nodes, and any of them up to q_h times. Any of these nodes contributes to the regret with a maximum amount of $BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^h$. Summing over all levels, we are sure that the expected cumulative regret cannot exceed this number. (3.25) follows from the definition of $v_2\rho^h$-packing number and the fact that any two nodes in $\mathcal{X}_h$ are at least $v_2\rho^h$

apart; in (3.26) we use Lemma 2.1, since $v_2\rho^h \leq v_2$, for any $h \geq 0$, and the fact that $q_h \leq 2\log(Tv_3)/(v_1\rho^h)^2$ in ACC-UCB.

Next, we bound $R_\alpha^2(T)$.

$$\begin{aligned} R_\alpha^2(T) &:= \sum_{\substack{t\leq T:\\ h(t)\geq H}} BK(6Nv_1/v_2+2)(v_1/v_2)v_2\rho^{h(t)} \\ &\leq TKB(6Nv_1/v_2+2)v_1\rho^H \ . \end{aligned} \tag{3.31}$$

From (3.30) and (3.31), we define

$$C_1 := 2QBK(6Nv_1/v_2+2)\frac{{v_2}^{-D_1}}{v_1(\rho^{-1}-1)}$$

and

$$C_2 := KB(6Nv_1/v_2+2)v_1 \ .$$

We have that

$$R_\alpha(T) \leq C_1\rho^{-H(D_1+1)}\log(Tv_3) + C_2\rho^H T \ .$$

If we let

$$H = \frac{\log T - \log(\log(Tv_3))}{(D_1+2)\log(1/\rho)}$$

so that (equivalently) we have

$$\begin{aligned} H &= \frac{\log T - \log(\log(Tv_3))}{(D_1+2)\log(1/\rho)} = \frac{\log\left(\frac{T}{\log(Tv_3)}\right)}{(D_1+2)\log(1/\rho)} = \frac{1}{(D_1+2)}\log_{1/\rho}\left(\frac{T}{\log(Tv_3)}\right) \\ &= -\log_\rho\left(\frac{T}{\log(Tv_3)}\right)^{\frac{1}{D_1+2}} \end{aligned}$$

and then we would obtain

$$\rho^{-H} = \left(\frac{T}{\log\left(Tv_3\right)}\right)^{\frac{1}{D_1+2}} \text{ and } \rho^{H} = \left(\frac{\log\left(Tv_3\right)}{T}\right)^{\frac{1}{D_1+2}} .$$

And thus, substituting in the bounds we get:

$$\begin{aligned}
R_\alpha(T) &\leq C_1\rho^{-H(D_1+1)}\log(Tv_3) + C_2\rho^H T \\
&= C_1 \cdot \left(\frac{T}{\log(Tv_3)}\right)^{\frac{D_1+1}{D_1+2}} \cdot \log(Tv_3) + C_2 \cdot \left(\frac{\log(Tv_3)}{T}\right)^{\frac{1}{D_1+2}} \cdot T \\
&= C_1 \cdot T^{1-\frac{1}{D_1+2}} \cdot (\log(Tv_3))^{1-\frac{D_1+1}{D_1+2}} + C_2 \cdot T^{1-\frac{1}{D_1+2}} \cdot (\log(Tv_3))^{\frac{1}{D_1+2}} \\
&= (C_1 + C_2) \cdot T^{1-\frac{1}{D_1+2}} \cdot (\log(Tv_3))^{\frac{1}{D_1+2}} .
\end{aligned}$$

■

3.2.2 Optimistic Regret Bounds

The regret bound in Theorem 3.1 depends on the worst super arm among the optimal super arms over all the rounds. While this being the worst among the best is reasonable, can we obtain a tighter bound? For this, we partition the set of the rounds $[T]$ into subsets and consider their contribution to the regret separately, in order to obtain different dimensions not depending on the worst of the best super arms overall, but the worst of the best super arms over smaller sets.

Formally, let $\pi : [T] \to [T]$ be a permutation of the rounds such that $u(f(\boldsymbol{x}_{\pi(1),S^{*\pi(1)}})) \geq \ldots \geq u(f(\boldsymbol{x}_{\pi(T),S^{*\pi(T)}}))$ and let us denote by $\pi([T])$ the new ordered set. Now let $\Gamma = \{\mathcal{T}_1^\Gamma, \ldots, \mathcal{T}_{|\Gamma|}^\Gamma\}$ be an ordered partition of $\pi([T])$ and let us denote by $T_\lambda^\Gamma = |\mathcal{T}_\lambda^\Gamma|$ the cardinality of $\mathcal{T}_\lambda^\Gamma$, for some $\lambda \in [|\Gamma|]$ (there are finitely many such partitions). Also, for any $0 < \lambda \leq |[\Gamma]|$, denote the expected reward of the

"worst" optimal super arm in $\mathcal{T}_\lambda^\Gamma$ as

$$u^*_{\min}(\Gamma,\lambda) = \min_{t\in\mathcal{T}_\lambda^\Gamma} u(f(\boldsymbol{x}_{t,S^{*t}}))$$

and the maximum cardinality of a super arm selected in one of the rounds in $\mathcal{T}_\lambda^\Gamma$ as

$$K(\Gamma,\lambda) = \max_{t\in\mathcal{T}_\lambda^\Gamma} K.$$

Let $\mathcal{P}(\pi([T]))$ be the set of all partitions of $\pi([T]))$. Now let us denote by $\bar{D}_\lambda^\Gamma$ the $(g_{\Gamma,\lambda}, u, \alpha u^*_{\min}(\Gamma,\lambda))$-optimality dimension associated with the subset $\mathcal{T}_\lambda^\Gamma$ and by $R_{\alpha,\lambda}(\mathcal{T}_\lambda^\Gamma)$ the regret incurred by the learner over the rounds in $\mathcal{T}_\lambda^\Gamma$.

Theorem 3.2. *Under the assumptions of Theorem 3.1, for any $\Gamma \in \mathcal{P}(\pi([T]))$ let $\{D_\lambda^\Gamma\}_{\lambda\le|\Gamma|}$ be a sequence of constants such that $D_\lambda^\Gamma > \bar{D}_\lambda^\Gamma$, for $\lambda \le |\Gamma|$. Then, there exists $Q_\lambda^\Gamma = Q_\lambda^\Gamma(\mathcal{X}, g_{\Gamma,\lambda}, u, f, \alpha u^*_{\min}(\Gamma,\lambda), c) > 0$, $\lambda \le |\Gamma|$ such that the α-regret incurred by ACC-UCB is upper bounded with probability at least $1 - 1/T$ as follows:*

$$R_\alpha(T) \le \sum_{\lambda=1}^{|\Gamma|} \left[\left(C_5(\Gamma,\lambda) + C_6(\Gamma,\lambda) \right) \cdot \left(T_\lambda^\Gamma\right)^{1-\frac{1}{D_\lambda^\Gamma+2}} \cdot \left(\log(Tv_3)\right)^{\frac{1}{D_\lambda^\Gamma+2}} \right],$$

where $C_5(\Gamma,\lambda) := 2Q_\lambda^\Gamma K(\Gamma,\lambda) B(6Nv_1/v_2 + 2)\frac{v_2^{-D_\lambda^\Gamma}}{v_1(1/\rho-1)}$ *and* $C_6(\Gamma,\lambda) := K(\Gamma,\lambda)B(6Nv_1/v_2+2)v_1$.

Proof. Fix an ordered partition $\Gamma \in \mathcal{P}(\pi([T]))$. Note that

$$\begin{aligned} R_\alpha(T) &= \sum_{\lambda=1}^{|\Gamma|} \left(\alpha \sum_{t\in\mathcal{T}_\lambda^\Gamma} \text{opt}(\boldsymbol{f}_t) - \sum_{t\in\mathcal{T}_\lambda^\Gamma} u(f(\boldsymbol{x}_{t,S_t})) \right) \\ &= \sum_{\lambda=1}^{|\Gamma|} R_{\alpha,\lambda}(\mathcal{T}_\lambda^\Gamma)\,, \end{aligned}$$

where $R_{\alpha,\lambda}(\mathcal{T}_\lambda^\Gamma)$ denotes the total regret incurred in round in $\mathcal{T}_\lambda^\Gamma$.

For any $\lambda \le |\Gamma|$, we can define a permutation $\zeta_\lambda^\Gamma : \mathcal{T}_\lambda^\Gamma \to [T_\lambda^\Gamma]$, such that $\zeta_\lambda(t_\lambda^{\Gamma_i}) = i$, for any $i \in [T_\lambda^\Gamma]$, where $t_\lambda^{\Gamma_i}$ is an element of the ordered set $\mathcal{T}_\lambda^\Gamma = \{t_\lambda^{\Gamma_1}, \ldots, t_\lambda^{\Gamma_\lambda}\}$. Thus,

abusing the terminology, we can solve the bandit subproblems with T_λ^Γ rounds, for each natural $\lambda \leq |\Gamma|$.

From Lemma 3.4 we have:

$$\begin{aligned} R_{\alpha,\lambda}(\mathcal{T}_\lambda^\Gamma) &\leq \sum_{t\in\mathcal{T}_\lambda^\Gamma} \sum_{k=1}^{K} B(6Nv_1/v_2+2)v_1\rho^{H_{t,k}} \\ &\leq \sum_{t\in\mathcal{T}_\lambda^\Gamma} BK(\Gamma,\lambda)(6Nv_1/v_2+2)v_1\rho^{h(t)} \ . \end{aligned}$$

We use an approach that is similar to the proof of Theorem 3.1. Fix a positive integer H_λ^Γ whose value will be determined later. Let

$$R_{\alpha,\lambda}^1(\mathcal{T}_\lambda^\Gamma) := \sum_{t\in\mathcal{T}_\lambda^\Gamma : h(t) < H_\lambda^\Gamma} BK(\Gamma,\lambda)(6Nv_1/v_2+2)v_1\rho^{h(t)}$$

and

$$R_{\alpha,\lambda}^2(\mathcal{T}_\lambda^\Gamma) := \sum_{t\in\mathcal{T}_\lambda^\Gamma : h(t) \geq H_\lambda^\Gamma} BK(\Gamma,\lambda)(6Nv_1/v_2+2)v_1\rho^{h(t)} \ .$$

We first bound $R_{\alpha,\lambda}^1(T_\lambda^\Gamma)$. Following steps analogous to (3.24)-(3.30) in the proof of Theorem 3.1, it can be shown that there exists some $Q_\lambda^\Gamma > 0$ (that depends only on $\mathcal{X}$,$g_{\Gamma,\lambda}$, u, f, $\alpha u_{\min}^*(\Gamma,\lambda)$ and c but not explicitly on T) such that for $D_\lambda^\Gamma > \bar{D}_\lambda^\Gamma$ we have

$$R_{\alpha,\lambda}^1(\mathcal{T}_\lambda^\Gamma) \leq C_5(\Gamma,\lambda)\cdot\rho^{-H_\lambda^\Gamma(D_\lambda^\Gamma+1)}\cdot\log(Tv_3) \ ,$$

where

$$C_5(\Gamma,\lambda) := 2Q_\lambda^\Gamma K(\Gamma,\lambda)B(6Nv_1/v_2+2)\frac{{v_2}^{-D_\lambda^\Gamma}}{v_1(1/\rho-1)} \ .$$

For $R_{\alpha,\lambda}^2(T_\lambda^\Gamma)$, we have

$$R_{\alpha,\lambda}^2(\mathcal{T}_\lambda^\Gamma) \leq C_6(\Gamma,\lambda)\cdot\rho^{H_\lambda^\Gamma}\cdot T_\lambda^\Gamma$$

where

$$C_6(\Gamma,\lambda) := K(\Gamma,\lambda)B(6Nv_1/v_2+2)v_1\ .$$

Now let us define H_λ^Γ analogously with Theorem 3.1, in order to obtain sublinear regret:

$$H_\lambda^\Gamma = \frac{\log T_\lambda^\Gamma - \log(\log(Tv_3))}{(D_\lambda^\Gamma+2)\log(1/\rho)}\ .$$

Substituting H_λ^Γ, we obtain the following.

$$\begin{aligned}R_{\alpha,\lambda}(\mathcal{T}_\lambda^\Gamma) &\le C_5(\Gamma,\lambda)\rho^{-H_\lambda^\Gamma(D_\lambda^\Gamma+1)}\log(Tv_3) + C_6(\Gamma,\lambda)\rho^{H_\lambda^\Gamma}T_\lambda^\Gamma \\ &\le C_5(\Gamma,\lambda)\cdot(T_\lambda^\Gamma)^{1-\frac{1}{D_\lambda^\Gamma+2}}\cdot(\log(Tv_3))^{\frac{1}{D_\lambda^\Gamma+2}} + C_6(\Gamma,\lambda)\cdot(T_\lambda^\Gamma)^{1-\frac{1}{D_\lambda^\Gamma+2}}\cdot(\log(Tv_3))^{\frac{1}{D_\lambda^\Gamma+2}}\ .\end{aligned}$$

Finally, by summing over all $\lambda \le |\Gamma|$, we obtain

$$\begin{aligned}R_\alpha(T) &= \sum_{\lambda=1}^{|\Gamma|} R_{\alpha,\lambda}(T_\lambda^\Gamma) \\ &\le \sum_{\lambda=1}^{|\Gamma|}\left[(C_5(\Gamma,\lambda)+C_6(\Gamma,\lambda))\cdot(T_\lambda^\Gamma)^{1-\frac{1}{D_\lambda^\Gamma+2}}\cdot(\log(Tv_3))^{\frac{1}{D_\lambda^\Gamma+2}}\right] \\ &\le \sum_{\lambda=1}^{|\Gamma|} O\left((T_\lambda^\Gamma)^{1-\frac{1}{D_\lambda^\Gamma+2}}\cdot(\log(Tv_3))^{\frac{1}{D_\lambda^\Gamma+2}}\right)\ .\end{aligned}$$

■

Corollary 3.1. *For $\xi \in (0,1)$ let $\bar{D}(\xi)$ be the $(g_\xi, u, \alpha u^*_{\min}(\xi))$-optimality dimension associated with the set $\mathcal{T}^\xi$. Let $D : (0,1) \to \mathbb{R}_+$ be any function such that $D(\xi) > \bar{D}(\xi)$, for all $\xi \in (0,1)$. Then, there exists $Q(\xi) = Q(\mathcal{X}, g_\xi, u, f, \alpha u^*_{\min}(\xi), c) > 0$ such that the α-regret incurred by ACC-UCB in T rounds is upper bounded with probability at least $1 - 1/T$ as follows:*

$$R_\alpha(T) \le \inf_{\xi\in(0,1)}\left((C_5(\xi)+C_6(\xi))\cdot T^{1-\frac{1}{D(\xi)+2}}\cdot(\log(Tv_3))^{\frac{1}{D(\xi)+2}} + \alpha u^*_{\max}T^\xi\right)$$

where $C_5(\xi) := 2Q(\xi)K(\xi)B(6Nv_1/v_2+2)\frac{{v_2}^{-D(\xi)}}{v_1(1/\rho-1)}$ *and* $C_6(\xi) = K(\xi)B(6Nv_1/v_2+2)v_1$.

Proof. Fix $\xi \in (0,1)$. Let $\mathcal{T}^\xi = \{\pi(1), \ldots, \pi(T - \lfloor T^\xi \rfloor)\}$, where π is the permutation defined in the beginning of the section. Also, let $W(\xi) = \max_{t\in\mathcal{T}^\xi} |S_t|$. The idea is to bound separately the regret coming from the rounds in $\mathcal{T}^\xi$ and those in $[T] - \mathcal{T}^\xi$. First we consider $\mathcal{T}^\xi$. From Lemma 3.4 we have:

$$\begin{aligned} R_\alpha(\mathcal{T}^\xi) &:= \alpha \sum_{t\in\mathcal{T}^\xi} \text{opt}(\boldsymbol{f}_t) - \sum_{t\in\mathcal{T}^\xi} u(f(\boldsymbol{x}_{t,S_t})) \\ &\leq \sum_{t\in\mathcal{T}^\xi} BK(\xi)(6Nv_1/v_2+2)v_2\rho^{h(t)} . \end{aligned}$$

Using the result of Theorem 3.2 the above argument can be bounded as

$$R_\alpha(\mathcal{T}^\xi) \leq (C_5(\xi) + C_6(\xi)) \cdot T^{1-\frac{1}{D(\xi)+2}} \cdot (\log(Tv_3))^{\frac{1}{D(\xi)+2}} ,$$

where

$$C_5(\xi) := 2Q(\xi)K(\xi)B(6Nv_1/v_2+2)\frac{{v_2}^{-D(\xi)}}{v_1(1/\rho-1)} ,$$

and

$$C_6(\xi) = K(\xi)B(6Nv_1/v_2+2)v_1 .$$

Here, $Q(\xi) > 0$ is a constant that depends on $\mathcal{X}$, u, f, $\alpha u^*_{\min}(\xi)$ and c but not explicitly on T. On the other hand, the regret incurred in round in $[T] - \mathcal{T}^\xi$ is upper bounded by $\alpha u^*_{\max} T^\xi$. The final result follows from summing these two bounds and taking the infimum. ■

3.3 Experiments

We consider a mobile crowdsourcing problem where the goal is to assign a subset of available workers to location dependent tasks arriving sequentially over time. Formally, in each round t, a task arrives with a location (normalized longitude and latitude) sampled uniformly at random from $[0, 1]^2$. Then, the learner selects $K \in \{2, 4\}$ workers from the set of available workers, $\mathcal{M}^t$, which is sampled from the Poisson distribution with mean 50. A worker is characterized by its location that lies in $[0, 1]^2$ and its energy and willingness to work (i.e. battery status) sampled uniformly at random from $[0, 1]$. We generate workers using the Gowalla dataset [47]. This dataset consists of 6,442,892 user check-ins in the social networking platform Gowalla. Each check-in contains the user id, time of check-in and location of check-in. In each round t, we randomly select M^t of these check-ins without replacement and normalize and assign each of their locations to a worker. Each base arm m in round t is a task-worker pair with a two-dimensional context $x_m^t = (x_{m,1}^t, x_{m,2}^t)$. Here, $x_{m,1}^t$ represents the normalized[3] Euclidean distance between the worker and task locations, while $x_{m,2}^t$ represents the battery status of the worker.

We define the expected base arm outcome as $f(x_m^t) = \bar{f}(x_{m,1}^t) \cdot (x_{m,2}^t)^2$ where $\bar{f}$ is a Gaussian probability density function with mean 0 and standard deviation 1. Note that f is decreasing in the distance between the worker and task and increasing in the worker's battery. Furthermore, the outcome $r(x_m^t)$ of worker m in round t is Bernoulli distributed with probability $f(x_m^t)$; it is 1 when the worker successfully completes the task and 0 otherwise. We assume that the task is successfully completed if at least one of the assigned workers completes the task, which is true for tasks such as cryptocurrency mining [48]. Hence, $u(r(\boldsymbol{x}_{S_t}^t)) = 1$ if $\sum_{m \in S^t} r(x_m^t) \geq 1$ and 0 otherwise.

We implemented the simulations in Python[4] and ran them for 50,000 rounds using ACC-UCB, CC-MAB [33] and random selections. For ACC-UCB, we set $v_1 = \sqrt{5}, v_2 = 1, \rho = 0.5$, and $N = 2$. Also, $x_{0,1}$ is a square with edge length 1 and center

[3]To normalize the distance we simply divide it by the maximum possible distance, $\sqrt{2}$.
[4]Full code is provided at https://github.com/Bilkent-CYBORG/ACC-UCB

$(0.5, 0.5)$. For CC-MAB we set $\alpha = 1$, $h_T = \left\lceil 50000^{\frac{1}{5}} \right\rceil$. We also used an exact oracle in both algorithms. Reported results correspond to averages over 10 independent runs. Figure 4.3 shows the cumulative regret and Figure 4.2 shows the average task reward up to t for all algorithms. As can be seen, ACC-UCB outperforms CC-MAB and random selections.

The time complexity of ACC-UCB can be analyzed as follows. At the beginning of round t, the learner observes the available contexts and identifies the available nodes, and thus the complexity for these operations is $|\mathcal{X}_t| \cdot |\mathcal{N}_t| \leq M^2$. After computing the indices of both nodes and contexts, which takes $|\mathcal{X}_t| + |\mathcal{N}_t| \leq 2M$, it feeds them to the oracle. Next, the algorithm gives $\boldsymbol{\theta}_t$ as input to the oracle in order to obtain S_t. In our experiments, since the oracle selects greedily, the time complexity for this operation is $O(|\mathcal{X}_t| \log K) \leq O(M \log K)$.[5] After this point we identify the selected nodes and for each such node $\chi_{h,i}$ we compute $\hat{\mu}_t(\chi_{h,i})$ and $C_t(\chi_{h,i})$, which takes $2 \cdot |\mathcal{P}_t| \leq 2K$. Overall, for T rounds, the time complexity is $O\left(\sum_{t=1}^{T}(M^2 + 2M + M \log K + 2K)\right) = O\left(((M^2 + 2M + M \log K + 2K) \cdot T\right)$.

[5]The oracle sorts all indices of contexts in $\mathcal{X}_t$ and then picks the top K of them.

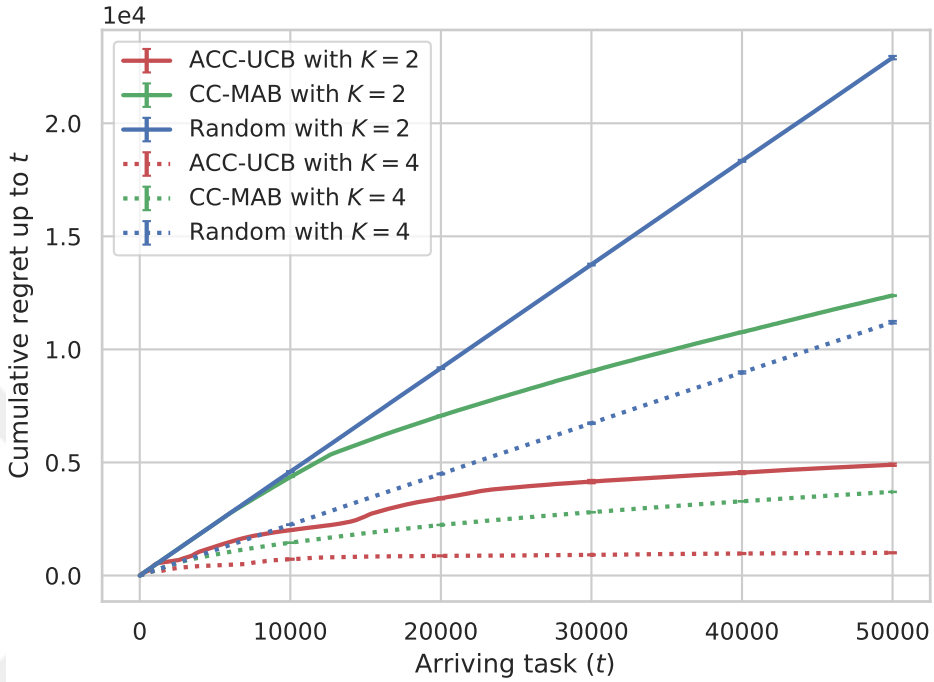

Figure 3.3: Cumulative regrets of algorithms.

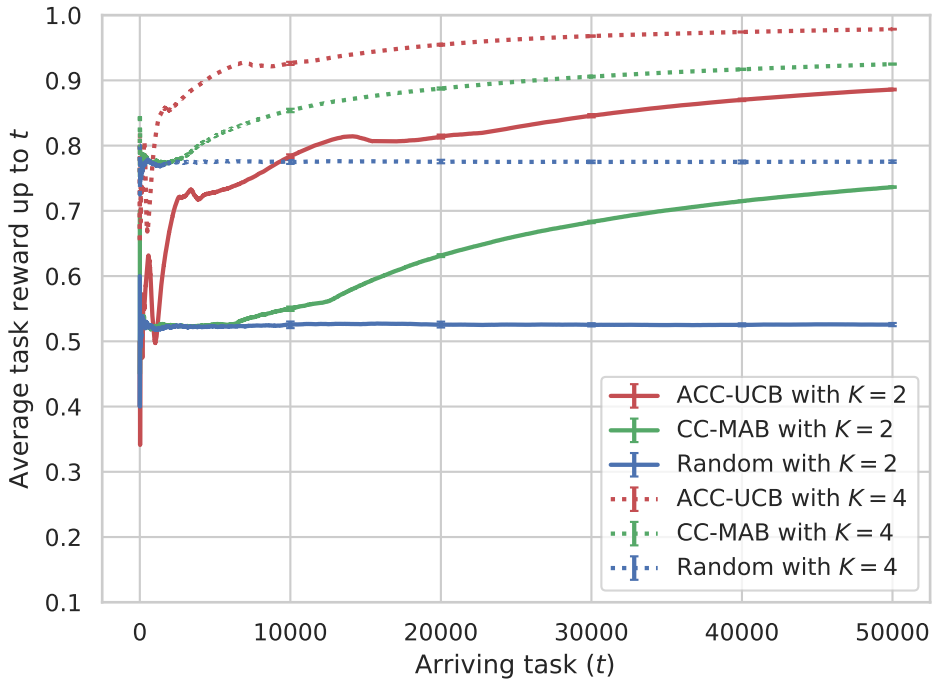

Figure 3.4: Average task reward up to round t.

Chapter 4

Solving CCV-MAB via Gaussian processes

4.1 The learning algorithm

Our second method is called *Optimistic Combinatorial Learning and Optimization with Kernel Upper Confidence Bounds* (O'CLOK-UCB), with pseudocode given in Algorithm 2. We assume that our latent function f is a sample from a GP as defined in Definition 2.3 and, furthermore, we assume that for every $x \in \mathcal{X}$, we have $k(x, x) \leq 1$. This is a general assumption that is widely used in the GP bandits literature [39]. The O'CLOK-UCB procedure is described as follows: At round t, we observe available base arms and their contexts. For each available base arm we maintain an index which is an upper confidence bound on its expected outcome. Let $S_1, \ldots, S_{t-1}$ be a sequence of super arms. Given base arm m with its associated context $x_{t,m}$, we define its GP-index (estimate of f using Gaussian processes) based on the observations $[\boldsymbol{r}_{S_1}^T, \ldots, \boldsymbol{r}_{S_{t-1}}^T]$ as:

$$I_t^{gp}(x_{t,m}) = \mu_{[\![t-1]\!]}(x_{t,m}) + \beta_t^{1/2}\sigma_{[\![t-1]\!]}(x_{t,m}) \,, \tag{4.1}$$

where $\mu_{[\![t-1]\!]}(\cdot)$ and $\sigma^2_{[\![t-1]\!]}(\cdot)$ stand for the posterior mean and variance (respectively) given the vector $[\boldsymbol{r}_{S_1}^T, \ldots, \boldsymbol{r}_{S_{t-1}}^T]$ of observations up to round t. Furthermore, β_t is a parameter that depends on t whose exact value will be specified later. We denote by $I_t^{gp}(\boldsymbol{x}_{t,S_t})$ the vector $[I_t^{gp}(x_{t,s_{t,1}}), \ldots, I_t^{gp}(x_{t,s_{t,|S_t|}})]^T$. Note that our definition of the GP-index of an arm with context x in the beginning of round t depends on $\mu_{[\![t-1]\!]}(x)$ and $\sigma_{[\![t-1]\!]}(x)$. This is due to the combinatorial nature of the problem. Since, at the beginning of round t, the algorithm has obtained semi-bandit feedback from $\sum_{t'=1}^{t-1} |S_{t'}|$ context selections, the posterior mean and variance of f calculated at that round will depend on the selected super arms, namely, $S_1, \ldots, S_{t-1}$.

Algorithm 2 O'CLOK-UCB

Input: $\mathcal{X}$, K, M; GP prior: $\mu_0 = \mu$, $k_0 = k$.
  Initialize: $\mu_0 = \mu$, $k_0 = k$.
  for $t = 1, \ldots, T$ **do**
    Observe base arms in $\mathcal{M}_t$ and their contexts $\mathcal{X}_t$.
    for $x_{t,m} : m \in \mathcal{M}_t$ **do**
      Calculate $\mu_{[\![t-1]\!]}(x_{t,m})$ and $\sigma_{[\![t-1]\!]}(x_{t,m})$ as in (2.1) and (2.3).
    end for
    Compute indices $I_t^{gp}(x_{t,m})$, $m \in \mathcal{M}_t$ as in (4.1).
    $S_t \leftarrow$ Oracle$(I_t^{gp}(x_{t,1}), \ldots, I_t^{gp}(x_{t,M_t}))$.
    Observe outcomes of base arms in S_t and collect the reward.
  end for

Similar to the other approach, after the indices of the available base arms (i.e., $\{I_t^{gp}(x_{t,m})\}_{m \in \mathcal{M}_t}$) are computed, they are given as input $\boldsymbol{\theta}_t$ to the approximation oracle in round t to obtain the super arm $S_t = (s_{t,1}, \ldots, s_{t,|S_t|}) \subseteq \mathcal{M}_t$ that will be played in round t. For this method, we assume a deterministic oracle in order to guarantee our theoretical results.

After observing the semi bandit feedback from the selected arms, at the beginning of the next round, we update the posterior distribution of f based on S_t according to rules (2.1) and (2.3), which will be used to compute the indices for the next round. Figure 4.1 illustrates the steps of our algorithm for round t.

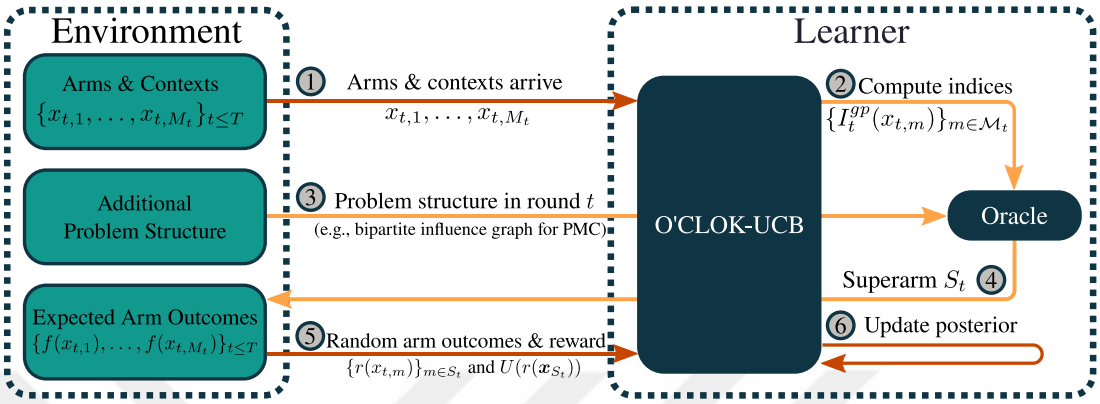

Figure 4.1: Illustration of the steps of our algorithm for round t. PMC stands for probabilistic maximum coverage.

4.2 Regret bounds

First, let us denote by $\boldsymbol{r}_{[\![t-1]\!]}$ the vector of observations made until the beginning of round t, that is,

$$\boldsymbol{r}_{[\![t-1]\!]} = [r^T(\boldsymbol{x}_{1,S_1}), \ldots, r^T(\boldsymbol{x}_{t-1,S_{t-1}})]^T \ .$$

For any $t \geq 1$, note that the posterior distribution of $f(x)$ given the observation vector $\boldsymbol{r}_{[\![t-1]\!]}$ is $\mathcal{N}(\mu_{[\![t-1]\!]}(x), \sigma^2_{[\![t-1]\!]}(x))$, for any $x \in \mathcal{X}_t$. Thus, applying the Gaussian tail bound, we obtain

$$\mathbb{P}\Big(|f(x) - \mu_{[\![t-1]\!]}(x)| > \beta_t^{1/2} \sigma_{[\![t-1]\!]}(x) \big| \boldsymbol{r}_{[\![t-1]\!]}\Big) \leq 2 \exp\left(\frac{-\beta_t}{2}\right) \ , \tag{4.2}$$

for $\beta_t \geq 0$. We will use this argument in order to prove our first result which gives high probability upper bounds on the deviation from the function value of the GP-index of any available context.

Our first result guarantees that the indices upper bound the expected base arm outcomes with high probability.

Lemma 4.1. *For any $\delta \in (0, 1)$, the probability of the following event $\mathcal{F}$ is at least*

$1-\delta$*:*

$$\{\forall t \geq 1, \forall x \in \mathcal{X}_t : |f(x) - \mu_{[\![t-1]\!]}(x)| \leq \beta_t^{1/2} \sigma_{[\![t-1]\!]}(x)\} \,,$$

where $\beta_t = 2\log(M\pi^2 t^2/3\delta)$.

Proof. We have:

$$\begin{aligned}
1 - \mathbb{P}(\mathcal{F}) &= \mathbb{E}\left[\mathbb{I}\left(\exists t \geq 1, \exists x \in \mathcal{X}_t : |f(x) - \mu_{[\![t-1]\!]}(x)| > \beta_t^{1/2}\sigma_{[\![t-1]\!]}(x)\right)\right] && (4.3)\\
&\leq \mathbb{E}\left[\sum_{t\geq 1}\sum_{x\in\mathcal{X}_t} \mathbb{I}\left(|f(x) - \mu_{[\![t-1]\!]}(x)| > \beta_t^{1/2}\sigma_{[\![t-1]\!]}(x)\right)\right] \\
&= \sum_{t\geq 1}\sum_{x\in\mathcal{X}_t} \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left(|f(x) - \mu_{[\![t-1]\!]}(x)| > \beta_t^{1/2}\sigma_{[\![t-1]\!]}(x)\right)\middle| \boldsymbol{r}_{[\![t-1]\!]}\right]\right] && (4.4)\\
&= \sum_{t\geq 1}\sum_{x\in\mathcal{X}_t} \mathbb{E}\left[\mathbb{P}\left(|f(x) - \mu_{[\![t-1]\!]}(x)| > \beta_t^{1/2}\sigma_{[\![t-1]\!]}(x)\middle| \boldsymbol{r}_{[\![t-1]\!]}\right)\right] \\
&\leq \sum_{t\geq 1}\sum_{x\in\mathcal{X}_t} 2\exp\left(\frac{-\beta_t}{2}\right) && (4.5)\\
&= 2M\sum_{t\geq 1}\frac{6\delta}{2M\pi^2 t^2} \\
&= \delta\frac{6}{\pi^2}\sum_{t\geq 1} t^{-2} = \delta \,,
\end{aligned}$$

where (4.3) follows from the fact that $\mathbb{P}(\mathcal{F}) = \mathbb{E}[\mathbb{I}(\mathcal{F})]$; (4.4) follows from the tower rule and the fact that the sets $\mathcal{X}_t$, $t \geq 1$ are fixed (thus there is no randomness from there); (4.5) follows from (4.2) and the rest follows from substituting β_t and the fact that $\sum_{t\geq 1} t^{-2} = \pi^2/6$. ■

Next, we upper bound the gap of a selected super arm in terms of the gaps of individual arms.

Lemma 4.2. *Given round* $t \geq 1$*, let us denote by* $S_t^* = \{s_{t,1}^*, \ldots, s_{t,|S_t^*|}^*\}$ *the optimal super arm in round* t*. Then, the following holds under the event* $\mathcal{F}$*:*

$$\alpha u(f(\boldsymbol{x}_{t,S_t^*})) - u(f(\boldsymbol{x}_{t,S_t})) \leq 2B\beta_t^{1/2}\sum_{k=1}^{|S_t|} |\sigma_{[\![t-1]\!]}(y_{t,k})| \,.$$

Proof. Let us first define $G_t = \text{argmax}_{S \in \mathcal{S}_t} u(I_t^{gp}(\boldsymbol{x}_{t,S}))$. Given that event $\mathcal{F}$ holds, we have:

$$\alpha \cdot u(f(\boldsymbol{x}_{t,S_t^*})) - u(f(\boldsymbol{x}_{t,S_t})) \leq \alpha \cdot u(I_t^{gp}(\boldsymbol{x}_{t,S_t^*})) - u(f(\boldsymbol{x}_{t,S_t})) \tag{4.6}$$

$$\leq \alpha \cdot u(I_t^{gp}(\boldsymbol{x}_{t,G_t})) - u(f(\boldsymbol{x}_{t,S_t})) \tag{4.7}$$

$$\leq u(I_t^{gp}(\boldsymbol{x}_{t,S_t})) - u(f(\boldsymbol{x}_{t,S_t})) \tag{4.8}$$

$$\leq B \sum_{k=1}^{|S_t|} |I_t^{gp}(\overline{x}_{t,k}) - f(\overline{x}_{t,k})| \tag{4.9}$$

$$\leq B \sum_{k=1}^{|S_t|} |\mu_{[\![t-1]\!]}(\overline{x}_{t,k}) - f(\overline{x}_{t,k})| + B \sum_{k=1}^{|S_t|} |\beta_t^{1/2} \sigma_{[\![t-1]\!]}(\overline{x}_{t,k})| \tag{4.10}$$

$$\leq 2B\beta_t^{1/2} \sum_{k=1}^{|S_t|} |\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})| \,, \tag{4.11}$$

where (4.6) follows from monotonicity of u and the fact that $f(x_{t,s_{t,k}^*}) \leq I_t^{gp}(x_{t,s_{t,k}^*})$, for $k \leq |S_t^*|$, by Lemma 4.1; (4.7) follows from the definition of G_t; (4.8) holds since S_t is the super arm chosen by the α-approximation Oracle; (4.9) follows from the Lipschitz continuity of u: (4.10) follows from the definition of GP-index and the triangle inequality; for (4.11) we use Lemma 4.1. ■

Before proving our main result, we prove an important lemma which will help us establish information-type regret bounds for O'CLOK-UCB.

Lemma 4.3. *Let $\boldsymbol{z}_t := \boldsymbol{x}_{t,S_t}$ be the vector of selected contexts at time $t \geq 1$. Given $T \geq 1$, we have:*

$$I\left(r(\boldsymbol{z}_{[T]}); f(\boldsymbol{z}_{[T]})\right) \geq \frac{1}{2K} \sum_{t=1}^{T} \sum_{k=1}^{|S_t|} \left(1 + \sigma^{-2} \sigma^2_{[\![t-1]\!]}(y_{t,k})\right) \,,$$

where $\boldsymbol{z}_{[T]} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T]^T$ is the vector of all selected contexts until round T.

Proof. By definition, we have

$$I\left(r(\boldsymbol{z}_{[T]}); f(\boldsymbol{z}_{[T]})\right) = H\left(r(\boldsymbol{z}_{[T]})\right) - H\left(r(\boldsymbol{z}_{[T]}) \middle| f(\boldsymbol{z}_{[T]})\right)$$

$$= H\left(r(\boldsymbol{z}_T), r(\boldsymbol{z}_{[T-1]})\right) - H\left(r(\boldsymbol{z}_{[T]})\middle|f(\boldsymbol{z}_{[T]})\right)$$
$$= H\left(r(\boldsymbol{z}_T)\middle|r(\boldsymbol{z}_{[T-1]})\right) + H\left(r(\boldsymbol{z}_{[T-1]})\right) - H\left(r(\boldsymbol{z}_{[T]})\middle|f(\boldsymbol{z}_{[T]})\right) \ .$$

Reiterating inductively we obtain:

$$I\left(r(\boldsymbol{z}_{[T]}); f(\boldsymbol{z}_{[T]})\right) = \sum_{t=2}^{T} H\left(r(\boldsymbol{z}_t)\middle|r(\boldsymbol{z}_{[t-1]})\right) + H\left(r(\boldsymbol{z}_1)\right) - H\left(r(\boldsymbol{z}_{[t]})\middle|f(\boldsymbol{z}_{[t]})\right) \ ,$$

since $r(\boldsymbol{z}_{[1]}) = r(\boldsymbol{z}_1)$. Let $\boldsymbol{\eta}_t = [\eta_{t,1}, \ldots, \eta_{t,|S_t|}]^T$. Note that

$$\begin{aligned} H\left(r(\boldsymbol{z}_{[t]})\middle|f(\boldsymbol{z}_{[t]})\right) &= H\left(f(\boldsymbol{z}_1) + \boldsymbol{\eta}_1, \ldots, f(\boldsymbol{z}_T) + \boldsymbol{\eta}_T\middle|f(\boldsymbol{z}_{[T]})\right) \\ &= H\left(f(\overline{x}_{1,1}) + \eta_{1,1}, \ldots, f(\overline{x}_{1,|S_1|}) + \eta_{1,|S_1|}, \ldots, f(\overline{x}_{T,1}) + \eta_{T,1}, \ldots,\right. \\ &\left. + f(\overline{x}_{T,|S_T|}) + \eta_{T,|S_T|}\middle|f(\boldsymbol{z}_{[T]})\right) \\ &= \sum_{t=1}^{T}\sum_{k=1}^{|S_t|} H(\eta_{t,k}) \\ &= \frac{1}{2}\sum_{t=1}^{T} \log\left|2\pi e\sigma^2 \boldsymbol{I}_{|S_t|}\right| \ , \end{aligned}$$

where the last equality follows from the entropy formula for Gaussian random variables. On the other hand, since any finite dimensional distributions associated with a Gaussian process are Gaussian random vectors and also, since $\boldsymbol{z}_t$ is deterministic given $\boldsymbol{z}_{[t-1]}$, the conditional distribution of $r(\boldsymbol{z}_t)$ given $r(\boldsymbol{z}_{[t-1]})$ is $\mathcal{N}\left(\mu_{[\![t-1]\!]}(\boldsymbol{z}_t), \Sigma_{[\![t-1]\!]}(\boldsymbol{z}_t) + \sigma^2\boldsymbol{I}_{|S_t|}\right)$, where $\mu_{[\![t-1]\!]}(\boldsymbol{z}_t) = [\mu_{[\![t-1]\!]}(\overline{x}_{t,1}), \ldots, \mu_{[\![t-1]\!]}(\overline{x}_{t,|S_t|})]^T$ and

$$\Sigma_{[\![t-1]\!]}(\boldsymbol{z}_t) = \begin{bmatrix} \sigma^2_{[\![t-1]\!]}(\overline{x}_{t,1}) & \ldots & k_{[\![t-1]\!]}(\overline{x}_{t,1}, \overline{x}_{t,|S_t|}) \\ \vdots & \ddots & \vdots \\ k_{[\![t-1]\!]}(\overline{x}_{t,|S_t|}, \overline{x}_{t,1}) & \ldots & \sigma^2_{[\![t-1]\!]}(\overline{x}_{t,|S_t|}) \end{bmatrix} \ .$$

Here, $k_0(x,y) = k(x,y)$ and $k_{[\![t-1]\!]}(x,y) := k_N(x,y)$, where $N = \sum_{t'=1}^{t-1}|S_{t'}|$.. Before we proceed, we need to emphasize the fact that CCGP-UCB does not employ any randomisation subroutines. There exist many deterministic α-approximation

oracles [34, 49, 50], including greedy approximation oracles. We assume that the α-approximation oracle that our algorithm uses is deterministic, and hence, there is no randomness coming from our algorithm. Therefore, we obtain the following.

$$
\begin{aligned}
I\left(r(\boldsymbol{z}_{[T]}); f(\boldsymbol{z}_{[T]})\right) &= \sum_{t=2}^{T} H\left(r(\boldsymbol{z}_t)\middle|r(\boldsymbol{z}_{[t-1]})\right) + H\left(r(\boldsymbol{z}_1)\right) - H\left(r(\boldsymbol{z}_{[t]})\middle|f(\boldsymbol{z}_{[t]})\right) \\
&= \sum_{t=2}^{T} \frac{1}{2}\left[\log\left|2\pi e\left(\Sigma_{[\![t-1]\!]}(\boldsymbol{z}_t) + \sigma^2 \boldsymbol{I}_{|S_t|}\right)\right|\right] \\
&\quad + \frac{1}{2}\left[\log\left|2\pi e\left(\Sigma_0(\boldsymbol{z}_1) + \sigma^2 \boldsymbol{I}_{|S_1|}\right)\right|\right] - \frac{1}{2}\sum_{t=1}^{T} \log\left|2\pi e \sigma^2 \boldsymbol{I}_{|S_t|}\right| \qquad (4.12) \\
&= \sum_{t=1}^{T} \frac{1}{2}\log\left|2\pi e\left(\Sigma_{[\![t-1]\!]}(\boldsymbol{z}_t) + \sigma^2 \boldsymbol{I}_{|S_t|}\right)\right| - \frac{1}{2}\sum_{t=1}^{T} \log\left|2\pi e \sigma^2 \boldsymbol{I}_{|S_t|}\right| \\
&= \sum_{t=1}^{T} \frac{1}{2}\log\left|2\pi e \sigma^2\left(\sigma^{-2}\Sigma_{[\![t-1]\!]}(\boldsymbol{z}_t) + \boldsymbol{I}_{|S_t|}\right)\right| - \frac{1}{2}\sum_{t=1}^{T} \log\left|2\pi e \sigma^2 \boldsymbol{I}_{|S_t|}\right| \\
&= \sum_{t=1}^{T} \frac{1}{2}\log\left|2\pi e \sigma^2 \boldsymbol{I}_{|S_t|}\right| + \sum_{t=1}^{T} \frac{1}{2}\log\left|\left(\sigma^{-2}\Sigma_{[\![t-1]\!]}(\boldsymbol{z}_t) + \boldsymbol{I}_{|S_t|}\right)\right| \\
&\quad - \frac{1}{2}\sum_{t=1}^{T} \log\left|2\pi e \sigma^2 \boldsymbol{I}_{|S_t|}\right| \\
&= \sum_{t=1}^{T} \frac{1}{2}\log\left|\left(\sigma^{-2}\Sigma_{[\![t-1]\!]}(\boldsymbol{z}_t) + \boldsymbol{I}_{|S_t|}\right)\right| \qquad (4.13) \\
&\geq \sum_{t=1}^{T} \frac{1}{2}\max_{1\leq k\leq |S_t|}\left(\log\left(1 + \sigma^{-2}\sigma^2_{[\![t-1]\!]}(\overline{x}_{t,k})\right)\right) \qquad (4.14) \\
&\geq \sum_{t=1}^{T}\sum_{k=1}^{|S_t|} \frac{1}{2|S_t|}\log\left(1 + \sigma^{-2}\sigma^2_{[\![t-1]\!]}(\overline{x}_{t,k})\right) \\
&\geq \sum_{t=1}^{T}\sum_{k=1}^{|S_t|} \frac{1}{2K}\log\left(1 + \sigma^{-2}\sigma^2_{[\![t-1]\!]}(\overline{x}_{t,k})\right) \ ,
\end{aligned}
$$

where (4.12) follows from the formula of conditional entropy for Gaussian random

vectors and for (4.14) we make the following observation. Given a symmetric positive-definite $n \times n$-matrix A, we can write $|A + \boldsymbol{I}_n| = \prod_{\lambda_i \leq n}(\lambda_i + 1) \geq \lambda^* + 1$, since each term is greater than or equal to 1, where $(\lambda_i)_{i \leq n}$ is the sequence of eigenvalues of A with its maximum eigenvalue being λ^*. Then, from the variational characterization of the maximum eigenvalue we can write

$$\lambda^* = \max_{v \in \mathbb{R}^n : \|v\|_2 = 1} v^t A v \geq \mathbf{1}_i^T A \mathbf{1}_i = a_{ii} \, ,$$

for each $i \leq n$. Here, the vector $\mathbf{1}_j$ represents the ith basis vector of $\mathbb{R}^n$. ■

Finally, we are ready to prove our main result.

Theorem 4.1. *Let $\delta \in (0, 1)$, $T \in \mathbb{N}$. Given $\beta_t = 2\log(M\pi^2 t^2/3\delta)$, the regret incurred by O'CLOK-UCB in T rounds is upper bounded with probability at least $1 - \delta$ as follows:*

$$R_\alpha(T) \leq K\sqrt{C\beta_T T \overline{\gamma}_T} \, ,$$

where $C = 8B^2/\log(1 + \sigma^{-2})$.

Proof. From Lemma 4.2 we have:

$$\begin{aligned} R_\alpha(T) &= \alpha \sum_{t=1}^{T} \text{opt}(f_t) - \sum_{t=1}^{T} u(f(\boldsymbol{x}_{t,S_t})) \\ &\leq 2B\beta_T^{1/2} \sum_{t=1}^{T} \sum_{k=1}^{|S_t|} |\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})| \, , \end{aligned} \tag{4.15}$$

using the fact that $\beta_t^{1/2}$ is monotonically increasing in t. To proceed, let us define a constant $L = \sigma^{-2}/\log(1 + \sigma^{-2})$. Taking the square of both sides, we have:

$$R_\alpha^2(T) \leq 4B^2\beta_T \left(\sum_{t=1}^{T} \sum_{k=1}^{|S_t|} |\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})| \right)^2$$

$$\leq 4B^2\beta_T T \sum_{t=1}^{T} \left(\sum_{k=1}^{|S_t|} |\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})| \right)^2 \tag{4.16}$$

$$\leq 4B^2\beta_T T \sum_{t=1}^{T} |S_t| \sum_{k=1}^{|S_t|} \left(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})\right)^2 \tag{4.17}$$

$$\leq 4B^2\beta_T TK \sum_{t=1}^{T} \sum_{k=1}^{|S_t|} \left(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})\right)^2$$

$$\leq 4B^2\beta_T TK\sigma^2 \sum_{t=1}^{T} \sum_{k=1}^{|S_t|} \sigma^{-2} \left(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})\right)^2 \tag{4.18}$$

$$\leq 4B^2\beta_T TK\sigma^2 L \cdot \left(\sum_{t=1}^{T} \sum_{k=1}^{|S_t|} \log\left(1 + \sigma^{-2} \left(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})\right)^2\right) \right) \tag{4.19}$$

$$\leq 8B^2\beta_T TK^2\sigma^2 LI\left(r(\boldsymbol{z}_{[T]}); f(\boldsymbol{z}_{[T]})\right) \tag{4.20}$$

$$\leq CK^2\beta_T T\overline{\gamma}_T\,, \tag{4.21}$$

where for (4.16) and (4.17) we have used the Cauchy Shwarz inequality twice; in (4.18) we just multiply by σ^2 and σ^{-2}; for (4.19) we use the following argument. Note that for any $i \in [0, \sigma^{-2}]$ we have that $i \leq L\log(1+i)$. Moreover, by assumption we have $|\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})| \leq 1$, for all $t \geq 1$, for all $k \leq K$, and thus, $\sigma^{-2}(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k}))^2 \leq \sigma^{-2}$. Therefore, we have that $\sigma^{-2}(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k}))^2 \leq L\log(1+\sigma^{-2}(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k}))^2)$; (4.20) follows from Lemma 4.3 and for (4.21) we use the definition of $\overline{\gamma}_T$.

Taking the square root of both sides we obtain our desired result. ■

Next, we provide tight regret bounds for the case when any feasible super arm S contains a fixed number of K base arms. However, it is worth mentioning that this restriction does not damage the practicality of the setting. Indeed, multi-play multi-armed bandits fall in this category. Another such prominent example is Crowdsourcing, which also features in our experiments. Furthermore, in this case, we prove lower and upper bounds on $\overline{\gamma}_T$. The following result gives a tighter bound than that of Theorem 4.1 which depends on the classical notion of maximum information gain.

Before continuing with the result, we state an auxiliary fact (for a proof see [51]) which will help us prove it.

Fact 1. *The predictive variance of a given context is monotonically non-increasing (i.e., given $x \in \mathcal{X}$ and the vector of selected samples $[x_1, \ldots, x_{N+1}]$, we have $\sigma^2_{N+1}(x) \leq \sigma^2_N(x)$, for any $N \geq 1$).*

Theorem 4.2. *Let $T, K \in \mathbb{N}$ and $\delta \in (0, 1)$. Under the conditions of Theorem 4.1 and the additional assumption that any feasible super arm has cardinality K, the regret incurred by O'CLOK-UCB is upper bounded as follows with probability at least $1 - \delta$:*

$$R_\alpha(T) \leq K\sqrt{C\beta_T T\gamma_T}\,,$$

where γ_T is as defined in Section 2.3 and C is as defined in Theorem 4.1.

Proof. Note that, in this case, the vector $\boldsymbol{z}_{[T]} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T]$ is composed of selected K-dimensional vectors of contexts in T rounds. Moreover, let us denote by $\boldsymbol{x}_{[T],k} = [\overline{x}_{1,k}, \ldots, \overline{x}_{T,k}]$ the subsequence of kth contexts for each $\boldsymbol{z}_t, t \leq T$ and for a given $x \in \mathcal{X}$ we denote by $\sigma^2_{[t],k}(x)$ the posterior variance of x given the locations of $\boldsymbol{x}_{[t],k}$. Furthermore, let $L = \sigma^{-2}/\log(1 + \sigma^{-2})$. Following up from (4.16) in the proof of Theorem 4.1, and noting that $|S_t| = K$ for all $t \in [T]$, we get:

$$\begin{aligned}
R^2_\alpha(T) &\leq 4B^2\beta_T \left(\sum_{t=1}^{T}\sum_{k=1}^{K} |\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})|\right)^2 \\
&\leq 4B^2\beta_T TK\sigma^2 L \cdot \left(\sum_{t=1}^{T}\sum_{k=1}^{K} \log\left(1 + \sigma^{-2}\left(\sigma_{[\![t-1]\!]}(\overline{x}_{t,k})\right)^2\right)\right) && (4.22)\\
&\leq 8B^2\beta_T TK\sigma^2 L \cdot \sum_{t=1}^{T}\sum_{k=1}^{K} \frac{1}{2}\log\left(1 + \sigma^{-2}\sigma^2_{[t-1],k}(\overline{x}_{t,k})\right) && (4.23)\\
&= 8B^2\beta_T TK\sigma^2 L \cdot \sum_{k=1}^{K}\left(\frac{1}{2}\sum_{t=1}^{T} \log\left(1 + \sigma^{-2}\sigma^2_{[t-1],k}(\overline{x}_{t,k})\right)\right) \\
&= 8B^2\beta_T TK\sigma^2 L \cdot \sum_{k=1}^{K} I\left(r(\boldsymbol{x}_{[T],k}); f(\boldsymbol{x}_{[T],k})\right) && (4.24)\\
&\leq 8B^2\beta_T TK^2\sigma^2 L \cdot \max_{\tilde{\boldsymbol{x}}_{[T]}} I\left(r(\tilde{\boldsymbol{x}}_{[T]}); f(\tilde{\boldsymbol{x}}_{[T]})\right) \\
&= 8B^2\beta_T TK^2\sigma^2 L \cdot \gamma_T \\
&= CK^2\beta_T T\gamma_T\,, && (4.25)
\end{aligned}$$

where for (4.22) we have used Cauchy-Schawrz twice and applied the same arguments as in the proof of Theorem 4.1; (4.23) follows from Lemma 1, where we note that the smaller the set of observations, the larger the predictive variance based on it, that is, the posterior variance of a given $x \in \mathcal{X}$ based on $\boldsymbol{z}_{[t-1]}$ is smaller than that based on $\boldsymbol{x}_{[t-1],k}$ since the latter is a subsequence of the former; (4.24) follows from Lemma 5.3 of [39].

Taking the square root of both sides we obtain our desired result. ■

We now state lower and upper bounds on $\overline{\gamma}_T$ in terms of the classical information gain. This gives us a measure of how tight our bounds are with respect to previous bounds.

Theorem 4.3. *Under the conditions of Theorem 4.2 we have $\overline{\gamma}_T \leq K\gamma_T$. Furthermore, in the non-volatile case*[1] *(i.e., when $\mathcal{X}_t = \mathcal{X}$, for all $t \geq 1$) we have*

$$\frac{1}{K}\gamma_{KT} \leq \overline{\gamma}_T \leq K\gamma_T \,,$$

where γ_{KT} is the classical maximum information gain over KT rounds.

Proof. Let $\tilde{\boldsymbol{z}}_{[T]} = [\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_T]$ be any given sequence of T K-dimensional vectors of contexts $x \in \mathcal{X}$ and let $\tilde{\boldsymbol{x}}_{[T],k} = [\tilde{x}_{1,k}, \ldots, \tilde{x}_{T,k}]$ be its subsequence of kth contexts for each $\tilde{\boldsymbol{z}}_t, t \leq T$. We have

$$\begin{aligned}
I\left(r(\tilde{\boldsymbol{z}}_{[T]}); f(\tilde{\boldsymbol{z}}_{[T]})\right) &= \sum_{t=1}^{T} \frac{1}{2} \log |\left(\sigma^{-2}\Sigma_{[\![t-1]\!]}(\tilde{\boldsymbol{z}}_t) + \boldsymbol{I}_K\right)| && (4.26)\\
&\leq \frac{1}{2}\sum_{t=1}^{T}\sum_{k=1}^{K} \log\left(1 + \sigma^{-2}\sigma^2_{[\![t-1]\!]}(\tilde{x}_{t,k})\right) && (4.27)\\
&\leq \sum_{k=1}^{K}\left(\frac{1}{2}\sum_{t=1}^{T} \log\left(1 + \sigma^{-2}\sigma^2_{[t-1],k}(\tilde{x}_{t,k})\right)\right) && (4.28)\\
&= \sum_{k=1}^{K} I\left(r(\tilde{\boldsymbol{x}}_{[T],k}); f(\tilde{\boldsymbol{x}}_{[T],k})\right) && (4.29)\\
&\leq K \max_{\tilde{\boldsymbol{x}}_{[T]}} I\left(r(\tilde{\boldsymbol{x}}_{[T]}); f(\tilde{\boldsymbol{x}}_{[T]})\right) && (4.30)
\end{aligned}$$

[1] Note that these bounds can be applied on the scenario when $|\mathcal{X}| < \infty$.

$$= K\gamma_T \ , \tag{4.31}$$

where (4.26) follows from (4.13); (4.27) follows as a consequence of Hadamard's inequality; (4.28) follows from Lemma 1 and finally, (4.29) follows from Lemma 5.3 of [39]. Since this inequality holds for any arbitrary sequence $\tilde{\boldsymbol{z}}_{[T]}$, then it also holds for the sequence that yields $\overline{\gamma}_T$ (see (4)).

On the other hand, for the sequence $[\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,K}, \ldots, \tilde{x}_{t,1}, \ldots, \tilde{x}_{t,k}]$, where $k \leq K$ and $x \in \mathcal{X}$, let us denote by $\sigma^2_{K(t-1)+k}(x)$ the posterior variance of x based on $[\tilde{x}_{1,1}, \ldots, \tilde{x}_{1,K}, \ldots, \tilde{x}_{t,1}, \ldots, \tilde{x}_{t,k}]$. If the available set of contexts is time-invariant, then we have:

$$\begin{aligned}
\overline{\gamma}_T &= \max_{\tilde{\boldsymbol{z}}_{[T]}:\tilde{z}_t \in \mathcal{Z}_t, t \leq T} I(r(\tilde{\boldsymbol{z}}_{[T]}); f(\tilde{\boldsymbol{z}}_{[T]})) \\
&\geq \max_{\tilde{\boldsymbol{z}}_{[T]}:\tilde{z}_t \in \mathcal{Z}, t \leq T} \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{1}{2K} \log\left(1 + \sigma^{-2}\sigma^2_{[\![t-1]\!]}(\tilde{x}_{t,k})\right) && (4.32) \\
&\geq \max_{\tilde{\boldsymbol{z}}_{[T]}:\tilde{z}_t \in \mathcal{Z}, t \leq T} \sum_{t=1}^{T} \sum_{k=1}^{K} \frac{1}{2K} \log\left(1 + \sigma^{-2}\sigma^2_{K(t-1)+k}(\tilde{x}_{t,k})\right) && (4.33) \\
&= \max_{A \subset \mathcal{X}:|A|=KT} \frac{1}{K} \left(\frac{1}{2} \sum_{s=1}^{KT} \log\left(1 + \sigma^{-2}\sigma^2_s(\tilde{x}_s)\right) \right) && (4.34) \\
&= \frac{1}{K} \max_{A \subset \mathcal{X}:|A|=KT} \left(\frac{1}{2} \sum_{s=1}^{KT} \log\left(1 + \sigma^{-2}\sigma^2_s(\tilde{x}_s)\right) \right) \\
&= \frac{1}{K}\gamma_{KT} \ , && (4.35)
\end{aligned}$$

where (4.32) follows from Lemma 4.3; (4.33) follows from Lemma 1 and for (4.34) we simply rewrite the double summation as a single sum following the natural order; for (4.35) we again use Lemma 5.3 of [39]. ■

Finally, using kernel-dependent explicit bounds on γ_T given in [39, 52], we state a corollary of Theorem 4.2 which gives similar bounds on the α-regret incurred by O'CLOK-UCB.

Remark 3. *Consider the case when $K = 1$. As a consequence of Theorem 4.3 we have that $\bar{\gamma}_T \leq \gamma_T$ in the volatile case, and $\bar{\gamma}_T = \gamma_T$ in the non-volatile case. The*

intuition behind this result is as follows. When the available arm set in round t is a subset of the ground set, learning about the optimal arm in that round (constrained over the available ones) does not require information about the other, non-available, arms. Thus, using γ_T in the volatile case would be using 'irrelevant information', and this counts for looser regret bounds (since regret grows proportionally with $\sqrt{\gamma_T}$). In the volatile case, focusing only on information about available arms in every round is enough. In this way, we obtain tighter bounds by using $\bar{\gamma}_T$, while in the non-volatile case we recover the classical notion of γ_T.

Corollary 4.1. *Let $\delta \in (0,1)$, $T, K \in \mathbb{N}$ and let $\mathcal{X} \subset \mathbb{R}^D$ be compact and convex. Under the conditions of Theorem 4.2 and for the following kernels, the α-regret incurred by O'CLOK-UCB in T rounds is upper bounded (up to polylog factors) with probability at least $1-\delta$ as follows: For the linear kernel we have: $R_\alpha(T) \leq \tilde{O}\left(K\sqrt{DT}\right)$; for the RBF kernel we have: $R_\alpha(T) \leq \tilde{O}\left(K\sqrt{T}\right)$; for the Matérn kernel we have: $R_\alpha(T) \leq \tilde{O}\left(KT^{(D+\nu)/(D+2\nu)}\right)$, where $\nu > 1$ is the Matérn parameter.*

Proof. This is an immediate application of the explicit bounds on γ_T given in Theorem 5 of [39], to the bound we obtained in Theorem 2. For the Matérn kernel, we have used the tighter bounds from [52]. ■

4.3 Experimental results

We evaluate the performance of O'CLOK-UCB by comparing it with ACC-UCB [36] and AOM-MC [33], two UCB-based algorithms. We perform semi-synthetic crowdsourcing simulations using the Foursquare dataset. The Foursquare dataset [53] contains check-in data from New York City (227,428 check-ins) and Tokyo (573,703 check-ins) for a period of 10 months from April 2012 to February 2013. Each check-in comes with a location tag as well as the time of the check-in. In our simulations, we use the NYC dataset because its locations are more spread out than the TKY dataset.

4.3.1 Simulation Setup

We perform a crowdsourcing simulation where the goal is to assign workers to arriving tasks. In our simulation, 250 tasks arrive (i.e., $T = 250$), each with a location (normalized longitude and latitude) sampled uniformly at random from $[0, 1]^2$ and a difficulty rating, sampled from $[0, 1]$ (0 is most difficult). Similarly, each worker also has a 2D location also in $[0, 1]^2$, sampled from the NYC dataset, and a battery status, sampled from $[0, 1]$.

In round t, the agent observes the set of available workers, the expected number of which is sampled from a Poisson distribution with mean 100. The set is composed of workers whose distance (Euclidean norm) to the task is less than $\sqrt{0.5}$. Then, each worker-task pair is an arm, with a three dimensional context comprised of the scaled distance between the worker and task[2], the task's difficulty, and the worker's battery. We also set a fixed budget constraint of $K = 5$, meaning that 5 workers much be chosen for each task.

Then, given the task-worker joint context (i.e., arm context) x and by noting that x^1, x^2, and x^3 represent the worker-task distance, task difficulty, and worker battery, respectively, we define the expected outcome of each base arm as $f(x) = \mathbb{E}[r(x)] = Ag\left(x^1\right)\sqrt{x^2 \cdot x^3}$, where g is a Gaussian probability density function with mean 0 and standard deviation 0.4, and $A = \sqrt{2\pi \cdot 0.4^2}$ is the scaling constant. Note that g is decreasing in the worker-task distance and task difficulty; and increasing in the worker's battery. Then, the random quality of each worker with joint context x is defined as $r(x) = f(x) + \eta$, where η is zero mean noise with standard deviation 0.1.

Finally, in round t and given K chosen workers with joint task-worker context vector $\boldsymbol{x} = [x_1, \ldots, x_K]$, we define our reward as $U(r(\boldsymbol{x})) = \log\left(1 + \sum_{i=1}^{K} r(x_i)\right)$. Notice that the $\log$ term reflects the diminishing return on having multiple workers. Moreover, since $\log$ is an increasing function, the Oracle for this setup will be a greedy one that picks the arms whose contexts have the highest performance estimates.

[2]To scale the distance we divide it by $\sqrt{0.5}$ because that is the maximum possible distance.

4.3.2 Algorithms

O'CLOK-UCB : We use the GPflow library [54] for the sampling from and updating the Gaussian Process. We set $\delta = 0.3$ and use a squared exponential kernel.

SO'CLOK-UCB : In this variation of our algorithm, we use the sparse approximation to the GP posterior described in [55]. In this sparse approximation, instead of using all of the arm contexts up to round t to compute the posterior, a small s element subset of them is used, called the inducing points. By using a sparse approximation, the time complexity of the posterior updating procedure reduces from $O(K^3 + (2K)^3 + \ldots + (KT)^3) = O(K^3T^4)$ to $O(s^2KT^2)$.

To ensure optimal performance, we pick these inducing points uniformly at random from all of the contexts picked so far. In other words, in round t, we sample s contexts from $\{x_{1,1}, \ldots x_{1,K}, \ldots x_{t-1,K}\}$. We run simulations with three different inducing points: 10, 20, and 50. We also set $\delta = 0.3$ and use the squared exponential kernel.

ACC-UCB: We set $v_1 = \sqrt{3}, v_2 = 1, \rho = 0.5$, and $N = 8$, as given in Definition 1 of [36]. The initial (root) context cell, $X_{0,1}$, is a three dimensional unit hypercube centered at $(0.5, 0.5, 0.5)$.

CC-MAB: Since we have a three dimensional context space, 300 tasks (rounds), and an exact Oracle, we set $h_T = \left\lceil 300^{\frac{1}{3 \cdot 1 + 3}} \right\rceil = 3$, as given in Theorem 1 of [33]. Hence, each hypercube has a length of $1/h_T = \frac{1}{3}$.

Benchmark: The benchmark greedily picks the arms with the highest expected outcome.

4.3.3 Results

We ran 5 independent runs and averaged the reward, regret, and time taken over these runs. We present the standard deviations of the runs as error bars. Figure 4.2 shows the average task reward up to task t, divided by the benchmark reward; and Figure 4.3

shows the cumulative regret of each algorithm for different T. The two UCB-based algorithms perform around the same and are equal in performance with SO'CLOK-UCB with 10 inducing points. However, they perform substantially worse for larger number of inducing points. Interestingly, only 100 inducing points were enough to achieve very close performance to O'CLOK-UCB , which does not use any approximations for the posterior update. Finally, as expected, the smaller the number of inducing points, the worse the performance of SO'CLOK-UCB compared with the non-sparse O'CLOK-UCB algorithm. However, even with 20 inducing points, SO'CLOK-UCB is able to outperform both CC-MAB and ACC-UCB by around 30%.

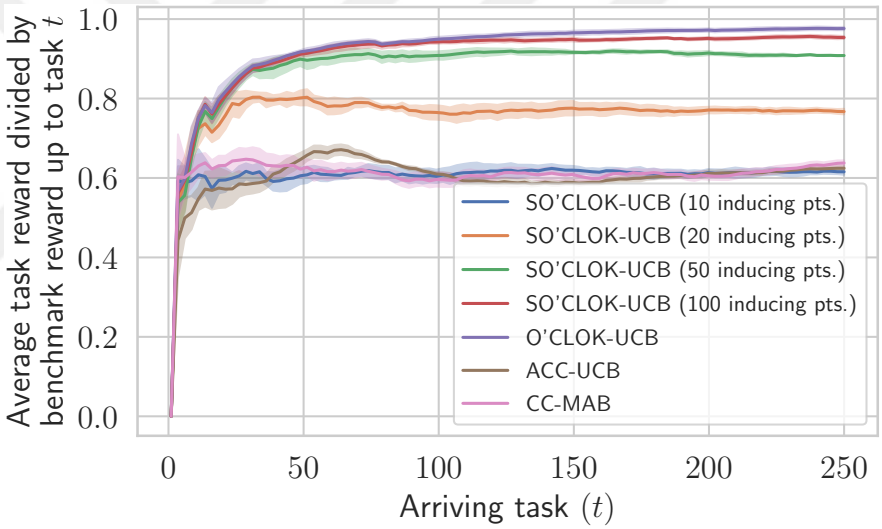

Figure 4.2: Average reward of each algorithm divided by that of the benchmark.

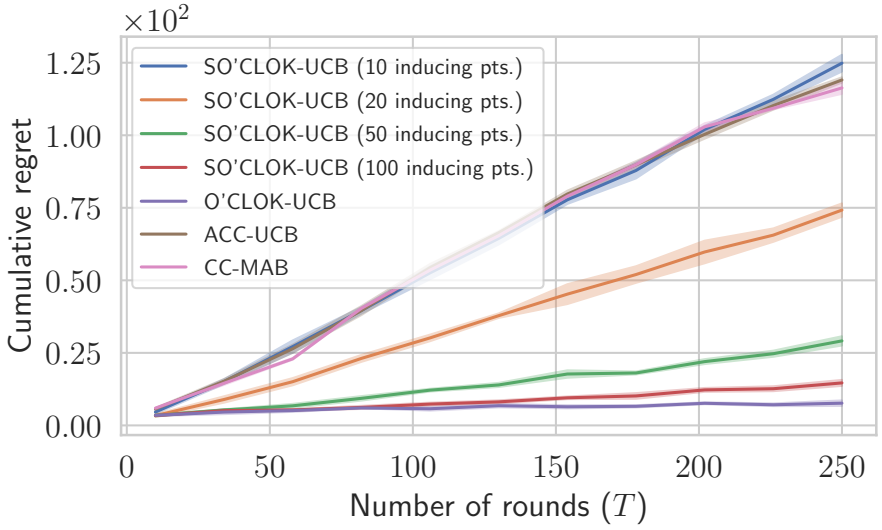

Figure 4.3: Cumulative regret of each algorithm for different number of rounds (T).

Chapter 5

Conclusion

We considered the contextual combinatorial volatile multi-armed bandit (CCV-MAB) problem with semi-bandit feedback, where in each round, the agent has to play a feasible subset of the base arms in order to maximize the cumulative reward. We proposed two online-learning methods to solve the CCV-MAB problem. Our first algorithm, called Adaptive Contextual Combinatorial Upper Confidence Bound (ACC-UCB), tradeoffs exploration and exploitation by performing adaptive discretization of the context space under mild continuity assumptions on the expected base arm outcomes and the expected reward. ACC-UCB is proven to achieve $\tilde{O}(T^{(\bar{D}+1)/(\bar{D}+2)+\epsilon})$ regret for any $\epsilon > 0$, where $\bar{D}$ represents the approximate optimality dimension associated with the context space. Our second algorithm, called Optimistic Combinatorial Learning and Optimization with Kernel Upper Confidence Bounds (O'CLOK-UCB) uses a Gaussian Process (GP) structure in order to estimate the true mean values of the base arm outcomes. Under the assumption that the expected base arm outcomes are drawn from a Gaussian process and that the expected reward is Lipschitz continuous with respect to the expected base arm outcomes, O'CLOK-UCB incurs $\tilde{O}(K\sqrt{T\bar{\gamma}_T})$ regret in T rounds, where $\bar{\gamma}_T$ is the maximum information gain with respect to the problem's constraints. In experiments, we showed that sparse GPs can be used to speed up UCB computation without significantly degrading the performance. Experimentally, both methods seem to substantially improve over the previous state-of-the-art. Furthermore O'CLOK-UCB outperforms ACC-UCB, both theoretically and experimentally.

Table 5.1: *Comparison with related works*. Below K represents the fixed cardinality of an feasible super arm; $\overline{D}$ is the approximate optimality dimension introduced in [36] and their bounds hold for any $\epsilon > 0$; η is the Hölder constant and ν is the Matérn parameter. The bounds are shown up to polylog factors.

Work	**Context space**	**Function**	**Smoothness assumption**	**Oracle (approx.)**	**Contextual & Volatile**	**Regret bounds**
[17]	Finite	Lipschitz	Explicit	(α, β)	No	$\tilde{O}\left(\sqrt{KT}\right)$
[33]	Infinite	Submodular	Explicit	$(1-1/e)$	Yes	$\tilde{O}\left(KT^{(D+2\eta)/(D+3\eta)}\right)$
[36] (ACC-UCB)	Compact	Lipschitz	Explicit	α	Yes	$\tilde{O}\left(KT^{(\overline{D}+1)/(\overline{D}+2)+\epsilon}\right)$
O'CLOCK-UCB (Linear)	Compact	Lipschitz	GP-induced	α	Yes	$\tilde{O}\left(K\sqrt{DT}\right)$
O'CLOCK-UCB (RBF)	Compact	Lipschitz	GP-induced	α	Yes	$\tilde{O}\left(K\sqrt{T}\right)$
O'CLOCK-UCB (Matérn)	Compact	Lipschitz	GP-induced	α	Yes	$\tilde{O}\left(KT^{(D+\nu)/(D+2\nu)}\right)$

In the table above we give a comparison of characteristics and regret bounds between O'CLOK-UCB and previous work in this setting, including ACC-UCB, in the case when the cardinality of feasible super arms is constant. Note that we achieve a substantial improvement from ACC-UCB and [33] on the order of the horizon in the case of linear and squared exponential kernels while maintaining an equal dependence on K, where K represents the budget. O'CLOK-UCB yields similar bounds to the classical combinatorial bandit [17] up to a $\sqrt{K}$ factor.

Table 5.2: Table of notation

Symbol	Meaning
K	The maximum possible cardinality of a feasible super arm
M	The maximum possible number of available arms in any given round $t \geq 1$
$(\mathcal{X}, d)$	The metric space of contexts (D-dimensional)
$\mathcal{M}_t$	The set of available base arms at round t
$\mathcal{S}_t$	The set of feasible super arms in round t
$x_{t,m}$	Context of base arm $m \in \mathcal{M}_t$ in round t
$\mathcal{X}_t$	$\{x_{t,m}\}_{w \in \mathcal{M}_t}$
$r(x_{t,m})$	(Random) outcome of base arm $m \in \mathcal{M}_t$ in round t
$f(x)$	Expected outcome of a base arm with context x
$\boldsymbol{f}_t$	$[f(x_{t,m})]_{m \in \mathcal{M}_t}$
$\boldsymbol{x}_{t,S}$	$[x_{t,s_1}, \ldots, x_{t,s_{\lvert S_t \rvert}}]$, $\lvert S_t \rvert$-tuple of contexts of base arms that are in super arm $S \in \mathcal{S}_t$
$u(\cdot)$	Reward function
opt$(\boldsymbol{f}_t)$	$\max_{S \in \mathcal{S}_t} u(f(\boldsymbol{x}_{t,S}))$
S_t	Super arm selected by the learner in round t
S_t^*	$\operatorname{argmax}_{S \in \mathcal{S}_t} u(f(\boldsymbol{x}_{t,S}))$, the optimal super arm in round t
$\chi_{h,i}$	The ith node in the hth level of the tree of partitions
$X_{h,i}$	The cell associated to node $\chi_{h,i}$
$\mathcal{X}_h$	$\{\chi_{h,i},\ 1 \leq i \leq N^h\}$
$M(\mathcal{X}, d, r)$	The r-packing number of $(\mathcal{X}, \lVert\cdot\rVert_2)$
$\mathcal{X}^{\kappa}_{g(r)}$	The $(g(r), u, \kappa)$-optimal set
$D^u(g, \kappa)$	The (g, u, κ)-optimality dimension
$I_t^{ad}(\chi_{h,i})$	The AD-index of node $\chi_{h,i}$ at time t
$I_t^{ad}(x_{t,m})$	The AD-index of base arm $m \in \mathcal{M}_t$ in round t
$I_t^{gp}(x)$	The GP-index of context x at time t
$\hat{\mu}_t(\chi_{h,i})$	The sample mean of the outcomes of the selected base arms with contexts in cell $X_{h,i}$ by round t
$C_t(\chi_{h,i})$	The number of times a base arm with context in $X_{h,i}$ was selected by round t
$\mu_{[\![t-1]\!]}(\cdot)$ and $\sigma^2_{[\![t-1]\!]}(\cdot)$	The posterior mean and variance of the GP
β_t	The posterior variance scaling parameter
$c_t(\chi_{h,i})$	The confidence radius of cell $X_{h,i}$ in round t
$(H_{t,k}, I_{t,k})$	(h, i) index of the active leaf node associated with the kth selected base arm ($s_{t,k}$) in round t
$(\tilde{h}_{t,m}, \tilde{i}_{t,m})$	(h, i) index of the active leaf node associated with base arm $m \in \mathcal{M}_t$ in round t
$\mathcal{L}_t$	The set of active leaf nodes in round t
$\mathcal{N}_t$	The set of available active leaf nodes in round t
$\mathcal{P}_t$	The set of active leaf nodes selected by the algorithm in round t

Bibliography

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933.

[2] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[4] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, no. 2-3, pp. 235–256, 2002.

[6] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," *arXiv preprint arXiv:1402.6028*, 2014.

[7] R. Agrawal, "Sample mean based index policies by o (log n) regret for the multi-armed bandit problem," *Adv. Appl. Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[8] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proc. 25th Annu. Conf. Learn. Theory*, pp. 39.1–39.26, 2012.

[9] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling," *Math. Oper. Res.*, vol. 39, no. 4, pp. 1221–1243, 2014.

[10] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 817–824, 2007.

[11] T. Lu, D. Pál, and M. Pál, "Contextual multi-armed bandits," in *Proc. 13th Int. Conf. Artif. Intell. and Statist.*, pp. 485–492, 2010.

[12] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proc. 14th Int. Conf. Artif. Intell. and Statist.*, pp. 208–214, 2011.

[13] A. Slivkins, "Contextual bandits with similarity information," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2533–2568, 2014.

[14] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. 19th Int. Conf. World Wide Web*, pp. 661–670, 2010.

[15] A. Tewari and S. A. Murphy, "From ads to interventions: Contextual bandits in mobile health," in *Mobile Health*, pp. 495–517, Springer International Publishing, 2017.

[16] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *J. Comput. Syst. Sci.*, vol. 78, no. 5, pp. 1404–1422, 2012.

[17] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *Proc. 30th Int. Conf. Mach. Learn.*, pp. 151–159, 2013.

[18] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu, "Combinatorial multi-armed bandit with general reward functions," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 1659–1667, 2016.

[19] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Proc. 18th Int. Conf. Artif. Intell. and Statist.*, pp. 535–543, 2015.

[20] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan, "Cascading bandits: Learning to rank in the cascade model," in *Proc. 32nd Int. Conf. Mach. Learn.*, pp. 767–776, 2015.

[21] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1466–1478, 2012.

[22] F. Radlinski, R. Kleinberg, and T. Joachims, "Learning diverse rankings with multi-armed bandits," in *Proc. 25th Int. Conf. Mach. Learn.*, pp. 784–791, 2008.

[23] P. Yang, N. Zhang, S. Zhang, K. Yang, L. Yu, and X. Shen, "Identifying the most valuable workers in fog-assisted spatial crowdsourcing," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1193–1203, 2017.

[24] W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1746–1778, 2016.

[25] S. Li, B. Wang, S. Zhang, and W. Chen, "Contextual combinatorial cascading bandits.," in *Proc. 33rd Int. Conf. Mach. Learn.*, vol. 16, pp. 1245–1253, 2016.

[26] Q. Wang and W. Chen, "Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications," *arXiv preprint arXiv:1703.01610*, 2017.

[27] A. Huyuk and C. Tekin, "Analysis of thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms," in *Proc. 22nd Int. Conf. Artif. Intell. and Statist.*, pp. 1322–1330, 2019.

[28] R. Kleinberg, A. Niculescu-Mizil, and Y. Sharma, "Regret bounds for sleeping experts and bandits," *Mach. Learn.*, vol. 80, no. 2-3, pp. 245–272, 2010.

[29] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1799–1813, 2019.

[30] Z. Bnaya, R. Puzis, R. Stern, and A. Felner, "Volatile multi-armed bandits for guaranteed targeted social crawling," in *Workshops at the Twenty-Seventh AAAI Conf. Artif. Intell.*, 2013.

[31] D. Chakrabarti, R. Kumar, F. Radlinski, and E. Upfal, "Mortal multi-armed bandits," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 273–280, 2009.

[32] A. Jain, A. D. Sarma, A. Parameswaran, and J. Widom, "Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace," *Proceedings of the VLDB Endowment*, vol. 10, no. 7, pp. 829–840, 2017.

[33] L. Chen, J. Xu, and Z. Lu, "Contextual combinatorial multi-armed bandits with volatile arms and submodular reward," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 3247–3256, 2018.

[34] L. Qin, S. Chen, and X. Zhu, "Contextual combinatorial bandit and its application on diversified online recommendation," in *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 461–469, SIAM, 2014.

[35] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári, "X-armed bandits," *J. Mach. Learn. Res.*, vol. 12, no. May, pp. 1655–1695, 2011.

[36] A. Nika, S. Elahi, and C. Tekin, "Contextual combinatorial volatile multi-armed bandit with adaptive discretization," in *Proc. 23rd Int. Conf. Artif. Intell. and Statist.*, vol. 108, pp. 1486–1496, 26–28 Aug 2020.

[37] S. Shekhar, T. Javidi, *et al.*, "Gaussian process bandits with adaptive discretization," *Electron. J. Stat.*, vol. 12, no. 2, pp. 3829–3874, 2018.

[38] E. Contal, D. Buffoni, A. Robicquet, and N. Vayatis, "Parallel gaussian process optimization with upper confidence bound and pure exploration," *Lecture Notes in Computer Science*, p. 225–240, 2013.

[39] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for gaussian process optimization in the bandit setting," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[40] M. A. Qureshi, A. Nika, and C. Tekin, "Multi-user small base station association via contextual combinatorial volatile bandits," *IEEE Transactions on Communications*, vol. 69, no. 6, pp. 3726–3740, 2021.

[41] O. Atan, W. R. Zame, and M. Schaar, "Sequential patient recruitment and allocation for adaptive clinical trials," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1891–1900, PMLR, 2019.

[42] L. A. Wolsey and G. L. Nemhauser, *Integer and combinatorial optimization*. John Wiley & Sons, 2014.

[43] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[44] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—i," *Mathematical programming*, vol. 14, no. 1, pp. 265–294, 1978.

[45] R. Munos, "Optimistic optimization of a deterministic function without the knowledge of its smoothness," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 783–791, 2011.

[46] J. L. Doob, *Stochastic processes*, vol. 101. New York Wiley, 1953.

[47] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery and Data Mining*, pp. 1082–1090, 2011.

[48] U. Mukhopadhyay, A. Skjellum, O. Hambolu, J. Oakley, L. Yu, and R. Brooks, "A brief survey of cryptocurrency systems," in *Proc. Int. Conf. Privacy, Security and Trust*, pp. 745–752, 12 2016.

[49] T. Lin, J. Li, and W. Chen, "Stochastic online greedy learning with semi-bandit feedbacks," in *Proc. Adv. Neural Inf. Process. Syst.*, pp. 352–360, 2015.

[50] N. Buchbinder and M. Feldman, "Deterministic algorithms for submodular maximization problems," *ACM Trans. Alg.*, vol. 14, no. 3, pp. 1–20, 2018.

[51] C. K. Williams and F. Vivarelli, "Upper and lower bounds on the learning curve for gaussian processes," *Machine Learning*, vol. 40, no. 1, pp. 77–102, 2000.

[52] S. Vakili, K. Khezeli, and V. Picheny, "On information gain and regret bounds in gaussian process bandits," *arXiv preprint arXiv:2009.06966*, 2020.

[53] D. Yang, D. Zhang, V. W. Zheng, and Z. Yu, "Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 1, pp. 129–142, 2015.

[54] A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman, " GPflow: A Gaussian process library using TensorFlow," *J. Mach. Learn. Res.*, vol. 18, pp. 1–6, apr 2017.

[55] M. K. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *Proc. 12th Int. Conf. Artif. Intell. and Statist.*, pp. 567–574, 2009.