Combinatorial Gaussian Process Bandits with Probabilistically Triggered Arms

İlker Demirel
Bilkent University, Ankara, Turkey

Cem Tekin
Bilkent University, Ankara, Turkey

Abstract

Combinatorial bandit models and algorithms are used in many sequential decision-making tasks ranging from item list recommendation to influence maximization. Typical algorithms proposed for combinatorial bandits, including combinatorial UCB (CUCB) and combinatorial Thompson sampling (CTS) do not exploit correlations between base arms during the learning process. Moreover, their regret is usually analyzed under independent base arm outcomes. In this paper, we use Gaussian Processes (GPs) to model correlations between base arms. In particular, we consider a combinatorial bandit model with probabilistically triggered arms, and assume that the expected base arm outcome function is a sample from a GP. We assume that the learner has access to an exact computation oracle, which returns an optimal solution given expected base arm outcomes, and analyze the regret of Combinatorial Gaussian Process Upper Confidence Bound (ComGP-UCB) algorithm for this setting. Under (triggering probability modulated) Lipschitz continuity assumption on the expected reward function, we derive $(O(\sqrt{mT\log T\gamma^{PTA}_{T,\boldsymbol{\mu}}}))$ $O(m\sqrt{\frac{T\log T}{p^*}})$ upper bounds for the regret of ComGP-UCB that hold with high probability, where m denotes the number of base arms, p^* denotes the minimum non-zero triggering probability, and $\gamma^{PTA}_{T,\boldsymbol{\mu}}$ denotes the pseudo-information gain. Finally, we show via simulations that when the correlations between base arm outcomes are strong, ComGP-UCB significantly outperforms CUCB and CTS.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 INTRODUCTION

Multi-armed bandits (MAB) are one of the simplest reinforcement learning models (Berry et al., 1985, Sutton et al., 1998, Robbins et al., 1952) in which the ultimate trade-off between exploration and exploitation can be elegantly analyzed. The simplest—yet most extensively studied—MAB model involves m arms whose expected outcome distributions are unknown. At each round, the learner selects one arm based on its past observations to maximize the (expected) cumulative reward. It only observes a noisy outcome from the selected arm at the end of each round, while other arms' outcomes remain hidden. A good learner should minimize its *regret*—the difference between the expected cumulative reward obtained by playing the optimal arms and that obtained by itself—by adapting its arm selection policy based on its history. When the arm outcomes are independent, it is well-known that any learner's regret will increase at least logarithmically over time (Lai et al., 1985), i.e., exploration is necessary and not free. Over decades, many different algorithms are proposed to minimize the regret, including but not limited to celebrated Thompson (posterior) sampling (Thompson et al., 1933, Agrawal et al., 2012, Russo et al., 2014) and upper confidence bound (UCB) (Lai et al., 1985, Agrawal et al., 1995, Auer et al., 2002) algorithms.

Unlike standard MAB, many sequential decision-making problems ranging from online influence maximization to bundle recommendation require multiple base arms to be simultaneously selected at each round. These—and many others—can be modeled as a combinatorial multi-armed bandit (CMAB), which is an extension of the original MAB problem, where at each round, the learner chooses a subset of base arms (a *super-arm*) (Gai et al., 2012, Kveton et al., 2015a, Chen et al., 2013, Gopalan et al., 2014), after which it observes the overall reward for the super-arm, and individual outcomes for each selected base arm. This type of feedback is called *semi-bandit feedback*. Authors of Kveton et al. (2015a) investigate the special case where the expected reward of a super-arm is the linear combination of its base arms' expected outcomes and derive

$O(Km \log T/\Delta)$ gap dependent and $O(\sqrt{KmT \log T})$ gap-free bounds for a combinatorial version of UCB1 in Auer et al. (2002). However, assuming a super-arm's expected reward to be a linear combination of expected base arm outcomes is a stringent assumption, which does not hold in many practical problems of interest. There are other works generalizing this setting to allow the expected reward of a super-arm to be a more general function of expected base arm outcomes (Chen et al., 2013).

In many real-world applications, the CMAB framework may fall short in terms of modeling the problem. For instance, the CMAB framework does not capture the probabilistic nature of the item list recommendation problem or influence maximization in a graph (Hüyük et al., 2020). For this reason, we are interested in a more general framework, called combinatorial multi-armed bandit with probabilistically triggered arms (CMAB-PTA) (Chen et al., 2016), where each base arm in a super-arm may trigger other arms depending on the triggering probabilities and base arm outcomes. The reward at the end of a round is a function of all triggered base arms and their expected outcomes. For this setup, it is shown that logarithmic in time regret is achievable when the expected reward function has ℓ_∞ bounded smoothness property (Chen et al., 2016). One drawback of this bound is that it depends on $1/p^*$, where p^* is the minimum triggering probability. In Wang et al. (2017), a bound free of $1/p^*$ is derived under more strict assumptions on the expected reward function, called triggering probability modulated (TPM) bounded smoothness. Moreover, Wang et al. (2017) shows that the dependence on $1/p^*$ is unavoidable in general. GPs offer a powerful and convenient framework for CMAB problems. The immediate advantage is that GPs can model the dependencies between different base arms' expected outcomes via a kernel function. Compared to other algorithms that assume independence between different base arm outcomes, a GP bandit uses every single base arm outcome observation to update its body of information about the other base arms. Besides, thanks to the analytical tractability and simple update equations of GPs, it is easy to implement and analyze such models.

Table 1: Our Work in Comparison to Related Work

ALGO.	PUBL.	Regret Bound
CUCB	(Chen et al., 2013)	$O(\sum_i \log T/\Delta_i)$
CUCB[1]	(Chen et al., 2016)	$O(\sum_i \log T/(p_i\Delta_i))$
CUCB[1]	(Wang et al., 2017)	$O(\sum_i \log T/\Delta_i)^*$
CTS[2]	(Wang et al., 2018)	$O(\sum_i \log T/\Delta_i)$
CTS[1,2]	(Hüyük et al., 2019)	$O(\sum_i \log T/(p_i\Delta_i))$
ComGP-UCB[1,2]		$O\left(m\sqrt{\frac{T\log T}{p^*}}\right)$
		$O\left(\sqrt{mT\log T\gamma^{PTA}_{T,\boldsymbol{\mu}}}\right)^*$

*With triggering probability modulated Lipschitz continuity
[1]PTA scenario considered
[2]Exact oracle used

In this work, we analyze the regret of ComGP-UCB, a Gaussian process upper confidence bound algorithm (Srinivas et al., 2010). We consider a CMAB-PTA framework where the learner has access to an exact computation oracle. We show that ComGP-UCB achieves $(O(\sqrt{mT\log T\gamma^{PTA}_{T,\boldsymbol{\mu}}}))$ $O(m\sqrt{T\log T/p^*})$ high probability regret bounds under (triggering probability modulated) Lipschitz continuity assumptions on the expected reward. Our bounds match the state-of-the-art in terms of dependence on time (Chen et al., 2013, Hüyük et al., 2020, Srinivas et al., 2010). Note that one of our bounds includes pseudo-information gain. This term is a function of the base arms' triggering probabilities and posterior variances and it captures the information that can not be captured by the algorithms which assume independence between base arm outcomes. We elaborate on the pseudo-information gain term, $\gamma^{PTA}_{T,\boldsymbol{\mu}}$, in Sections 4 and 5.

The rest of the paper is organized as follows. Section 2 formulates the CMAB-PTA problem. Section 3 describes the ComGP-UCB algorithm. Section 4 analyzes the regret of ComGP-UCB. Section 5 contains the numerical results, and Section 6 presents the concluding remarks. Proofs of the lemmas used and additional results are provided in the supplementary material.

2 PROBLEM FORMULATION

We consider a sequential decision making problem with time horizon T. The learner interacts with its environment through m base arms indexed by the set $[m] \coloneqq \{1, \dots, m\}$, over rounds indexed by $t \in [T]$. Every base arm $i \in [m]$ is characterized by its context vector $\boldsymbol{x}_i \in \mathcal{X}$, where $\mathcal{X} \coloneqq [0,1]^d$ is the d-dimensional context set. In each round, the following actions take place in order:

- The learner selects a super arm $S(t)$ from a subset of $2^{[m]}$, denoted by $\mathcal{I}$.
- The arms in $S(t)$ may trigger other arms via a stochastic triggering process, D^{trig}, yielding a set of base arms $S'(t)$ which is a super set of $S(t)$. We denote the set of arms that can be possibly triggered by the arms in S by $\tilde{S}$.
- The learner observes the individual outcomes of each base arm in $S'(t)$, and a reward that depends on $S'(t)$ (*semi-bandit feedback*).

2.1 Base Arm Outcomes

We model the expected base arm outcomes as a sample from a Gaussian process defined on $\mathcal{X}$ (Rasmussen et al., 2004). A Gaussian process $GP(m(\boldsymbol{x}),\ k(\boldsymbol{x},\boldsymbol{x}'))$ can be thought of as a probability distribution over functions, or an infinite-dimensional multivariate Gaussian distribution. It is characterized by its mean, $m(\boldsymbol{x})$, and kernel, $k(\boldsymbol{x},\boldsymbol{x}')$, functions. GPs enjoy non-parametric flexibility and analytical tractability properties. Using GPs to model the expected outcome function, we gain the power to exploit the dependencies between the expected outcomes of different base arms via the kernel function.

At each round t, the environment draws a random outcome vector $\boldsymbol{Y}(t) \coloneqq (Y_1(t),\dots,Y_m(t))$ from the true outcome distribution $\mathcal{D}$, which is unknown to the learner, where $Y_i(t)$ is the outcome of base arm i at round t. $Y_i(t) \coloneqq \mu_i + \epsilon_{i,t}$, where $\epsilon_{i,t}$ are i.i.d. zero mean Gaussian random variables with known variance σ^2, which are used in updating the posterior mean vector and covariance matrix of the GP. Let $\mu_i \coloneqq \mathbb{E}_{\boldsymbol{Y}\sim\mathcal{D}}[Y_i(t)]$ denote the expected outcome of base arm i, $\boldsymbol{\mu} \coloneqq (\mu_1,\dots,\mu_m)$ denote the expected base arm outcome vector, and $\boldsymbol{\mu}_S$ denote the projection of $\boldsymbol{\mu}$ onto S. We assume that there exists a function $f:\mathcal{X}\to\mathbb{R}$, sampled from a GP, such that $f(\boldsymbol{x}_i)=\mu_i$.

2.2 Triggering Process

In every round t, the learner selects a super arm $S(t)\in\mathcal{I}$ based on its past observations. All the arms in the selected super arm are immediately triggered. These arms can trigger other arms, and triggered arms can further trigger other arms, until no more arm is triggered. By the end of the triggering process, we end up with a set of trigered arms $S'(t)\sim D^{trig}(S(t),\boldsymbol{Y}(t))$, such that $S(t)\subseteq S'(t)\subseteq \tilde{S}(t)\subseteq [m]$. The triggering process can be described via a set of triggering probabilities. For every base arm $i\in[m]$, and $S\in\mathcal{I}$, $p_i^{\mathcal{D}',S}$ denotes the triggering probability of the base arm i when super arm S is selected and the base arm outcome distribution is $\mathcal{D}'$. We let $p_i^S \coloneqq p_i^{\mathcal{D},S}$, where $\mathcal{D}\in\mathbb{D}$ is the true arm outcome distribution, and $\mathbb{D}$ is the class of distributions that $\mathcal{D}$ belongs to. $\mathcal{D}$ is unknown by the learner, but $\mathbb{D}$ is known. In our case, $\mathbb{D}$ is the class of m-dimensional multivariate Gaussian distributions, since we assume that the expected base arm outcomes are sampled from a GP. We define $p_i \coloneqq \min\{p_i^S \mid S\in\mathcal{I}, p_i^S>0\}$ as the minimum non-zero triggering probability of the base arm i, where $i\in\{1,\dots,m\}$ and $p^* \coloneqq \min_{i\in[m]} p_i$ as the minimum non-zero triggering probability.

2.3 Reward

After each round t, the learner observes a reward, $R(S'(t),\boldsymbol{Y}(t))$, depending on the set of triggered arms and the true expected base arm outcome vector. To simplify the notation, we use $R(t)=R(S'(t),\boldsymbol{Y}(t))$.

Assumption 1. *Given $\boldsymbol{\mu}$, we assume that the expected reward of a super arm is only a function of the super arm and the true outcome vector, and denote $\mathbb{E}[R(t)|\boldsymbol{\mu}] = r(S(t),\boldsymbol{\mu})$.*

Now let us define the set of optimal super arms (i.e., the ones that maximize $R(t)$) for a parameter vector $\boldsymbol{\theta}$ as $OPT(\boldsymbol{\theta}) \coloneqq \arg\max_{S\in\mathcal{I}} r(S,\boldsymbol{\theta})$. OPT denotes the exact oracle, which is a problem specific computation black-box. It knows the problem structure (e.g., cascading bandit, influence maximization, probabilistic maximum coverage) and calculates the optimal super-arm for a given base arm outcome vector (i.e., $\boldsymbol{\theta}$). For instance, in an influence maximization problem, the oracle knows the graph structure, and it can provide the optimal solution given the triggering probabilities of the edges. Let $S^*\in OPT(\boldsymbol{\mu})$ represent an optimal super arm. Based on this information, the learner aims to minimize the cumulative regret over the time horizon T, which can be written as,

$$Reg_{\boldsymbol{\mu}}(T) \coloneqq \sum_{t=1}^{T} r(S^*,\boldsymbol{\mu}) - r(S(t),\boldsymbol{\mu})\ .$$

Next, we make some standard assumptions on the reward function R (Hüyük et al., 2020, Chen et al., 2016, Wang et al., 2018).

Assumption 2. *(Monotonicity) The expected reward, $r(S,\boldsymbol{\mu})$, is non-decreasing w.r.t. $\boldsymbol{\mu}$, that is, given $\boldsymbol{\mu}$, $\boldsymbol{\nu}\in\mathbb{R}^M$ such that $\mu_i\geq\nu_i$, $\forall i\in[M]$, we have $r(S,\boldsymbol{\mu})\geq r(S,\boldsymbol{\nu}), \forall S\in\mathcal{I}$.*

Assumption 3. *(Lipschitz continuity) There exists a constant $B>0$, such that for every super arm S and every pair of mean base arm outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, we have $|r(S,\boldsymbol{\mu})-r(S,\boldsymbol{\nu})|\leq B||\boldsymbol{\mu}_{\tilde{S}}-\boldsymbol{\nu}_{\tilde{S}}||_1$, where $||\cdot||_1$ denotes the ℓ_1 norm.*

Assumption 4. *(Triggering probability modulated Lipschitz continuity) There exists a constant $\tilde{B}>0$, such that for every super arm S and every pair of base arm outcome distributions $\mathcal{D}$ and $\mathcal{D}'$ with mean outcome vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ respectively, we have,*

$$|r(S,\boldsymbol{\mu})-r(S,\boldsymbol{\nu})| \leq \tilde{B}\sum_{i\in\tilde{S}} p_i^{\mathcal{D},S}|\mu_i-\nu_i|\ .$$

2.4 Observation Model

Since we assume semi-bandit feedback, the individual outcomes of each base arm in $S'(t)$ are revealed after each round. That is, $Q(S'(t),\boldsymbol{Y}(t)) \coloneqq \{(i,Y_i(t)\ :$

$i \in S'(t))\}$ is revealed. We let $Q(t) = Q(S'(t), \boldsymbol{Y}(t))$. By these definitions, the information available to the learner to guide its actions in round $t+1$ is $\mathcal{F}_t := \{(S(\tau), Q(\tau)) : \tau \in [t]\}$. After making a set of observations, we can update the posterior mean and variance at $\boldsymbol{x} \in \mathcal{X}$ as follows,

$$k_{N(t)}(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x},\ \boldsymbol{x}') - \boldsymbol{k}_{N(t)}(\boldsymbol{x})^T(\boldsymbol{K}_{N(t)} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_{N(t)}(\boldsymbol{x}') \quad (1)$$

$$\hat{\sigma}^2_{N(t)}(\boldsymbol{x}) = k_{N(t)}(\boldsymbol{x},\ \boldsymbol{x}) \quad (2)$$

$$\hat{\mu}_{N(t)}(\boldsymbol{x}) = \boldsymbol{k}_{N(t)}(\boldsymbol{x})^T(\boldsymbol{K}_{N(t)} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{Y}_t, \quad (3)$$

where $\boldsymbol{Y}_t := [\boldsymbol{Y}^T(1), \dots, \boldsymbol{Y}^T(t-1)]^T$ denotes the vector of observations made until round t, and $N(t) := |\boldsymbol{Y}_t|$ denotes the total number of observations made until round t. Moreover, $k_{N(t)}(\boldsymbol{x},\ \boldsymbol{x}')$ denotes the posterior covariance between $\boldsymbol{x}$ and $\boldsymbol{x}'$, and $\hat{\mu}_{N(t)}(\boldsymbol{x})$ and $\hat{\sigma}^2_{N(t)}(\boldsymbol{x})$ denote the posterior mean and variance at $\boldsymbol{x} \in \mathcal{X}$ at round t, respectively. $\boldsymbol{k}_{N(t)}(\boldsymbol{x}) := [k(\boldsymbol{x}^1,\ \boldsymbol{x}), \dots, k(\boldsymbol{x}^{N(t)},\ \boldsymbol{x})]^T$ denotes the vector of covariances between $\boldsymbol{x} \in \mathcal{X}$, and past observations $[\boldsymbol{x}^1, \dots, \boldsymbol{x}^{N(t)}]$, where $\boldsymbol{x}^i$ is the ith base arm picked from the beginning. $\boldsymbol{K}_{N(t)}$ is the gram matrix, $\boldsymbol{I}$ is the $N(t) \times N(t)$ identity matrix, and σ^2 is the noise variance that we include in our calculations.

To summarize, $([m], \mathcal{I}, \mathcal{D}, D^{trig}, R)$ tuple characterizes a CMAB-PTA problem instance, among which only $\mathcal{D}$ is unknown to the learner.

3 ComGP-UCB ALGORITHM

Pseudo-code for ComGP-UCB is given in Algorithm 1, and Figure 1 demonstrates the algorithm flow. ComGP-UCB uses Gaussian process based upper confidence bounds to form optimistic estimates of the expected base arm outcomes (Srinivas et al., 2010). First, it initiates a GP, g, in the base arm domain. At the beginning of each round, the algorithm calculates an upper confidence bound (UCB) for base arms as follows,

$$\bar{\mu}_t(\boldsymbol{x}_i) = \hat{\mu}_{N(t)}(\boldsymbol{x}_i) + \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i),$$

where $i \in [m]$, and $\hat{\mu}_{N(t)}$ and $\hat{\sigma}^2_{N(t)}$ denote the posterior mean and variance vectors of g at round t, respectively.

Algorithm 1 ComGP-UCB

1: **Initialize** $g \sim GP(\hat{\mu}_0(\boldsymbol{x}),\ k_0(\boldsymbol{x}, \boldsymbol{x}'))$
2: **for** $t = 1, \dots, T$ **do**
3: $\quad \bar{\boldsymbol{\mu}}_t = \{\hat{\mu}_{N(t)}(\boldsymbol{x}_i) + \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)\}_{i=1}^m$
4: $\quad S(t) = OPT(\bar{\boldsymbol{\mu}}_t)$
5: $\quad$ Play $S(t)$ and observe $Q(t)$
6: $\quad$ Update $\mathcal{F}_t$ based on $Q(t)$
7: $\quad$ Update $\hat{\mu}_{N(t+1)}(\boldsymbol{x}_i)$ and $\hat{\sigma}_{N(t+1)}(\boldsymbol{x}_i)$, $i \in [m]$
8: **end for**

Without loss of generality, we assume bounded variance, that is, $\hat{\sigma}^2_{N(t)}(x) = k_{N(t)}(\boldsymbol{x}, \boldsymbol{x}) \leq 1$ (Srinivas et al., 2010). $\beta_n := 2\log(\frac{mT\pi_n}{\delta})$ is the exploration coefficient where $\sum_{n=1}^{\infty} 1/\pi_n = 1$ (e.g., $\pi_n = \pi^2 n^2/6$). Note that β_n is strictly increasing in n. We define $\bar{\boldsymbol{\mu}}_t := [\bar{\mu}_t(\boldsymbol{x}_i), \dots, \bar{\mu}_t(\boldsymbol{x}_m)]$ to denote the vector of estimated optimistic expected base arm outcomes at round t. Then, $\bar{\boldsymbol{\mu}}_t$ is passed to the exact oracle to select the super-arm in round t, $S(t)$.[1] After $S(t)$ is selected, $Q(t)$ is observed and the mean vector and covariance matrix of g are updated according to (1) and (3).

4 REGRET ANALYSIS

4.1 Main results

In this section, we bound the regret of ComGP-UCB when the reward function satisfies (TPM) Lipschitz continuity. First, we derive an upper bound on the regret when the reward function satisfies Lipschitz continuity.

Theorem 1. *Given δ, $\rho \in (0,1)$, $\alpha \in [mT]$, and under Assumptions 1, 2, and 3, the cumulative regret of ComGP-UCB after round T is upper bounded with at least $1 - \delta - \frac{m}{\alpha} - \frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$ probability as follows,*

$$\mathbb{P}\Big\{Reg_{\boldsymbol{\mu}}(T) \leq 4mB\sqrt{\frac{T\beta_{mT}\sigma^2}{(1-\rho)p^*}} + 2m\alpha B\sqrt{\beta_{mT}}\Big\} \geq 1 - \delta - \frac{m}{\alpha} - \frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}],$$

where B is the Lipschitz constant.

We observe that the right hand-side in Theorem 1 is greater than zero when $\alpha > \frac{m + \frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]}{1-\delta}$. This condition is translated into a condition on T in the following corollary. Also, note that $\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$ should be finite in order for this bound to make sense. One simple example is the case when $p^* \geq p^*_{\min}$ for all $\boldsymbol{\mu}$ for a constant $p^*_{\min} > 0$.

Corollary 1. *Setting $\alpha = \sqrt{T}$ in Theorem 1, we have,*

$$\mathbb{P}\Big\{Reg_{\boldsymbol{\mu}}(T) \leq 4mB\sqrt{\frac{T\beta_{mT}\sigma^2}{(1-\rho)p^*}} + 2mB\sqrt{T\beta_{mT}}\Big\} \geq 1 - 2\delta,$$

when $\sqrt{T} > \frac{m + \frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]}{\delta}$.

Next, we bound the regret when the expected reward function satisfies the TPM Lipschitz continuity. The

[1] We note that our regret analysis can easily be extended to the cases when approximation oracles are used, by redefining the regret to be α-approximation regret (Chen et al., 2013).

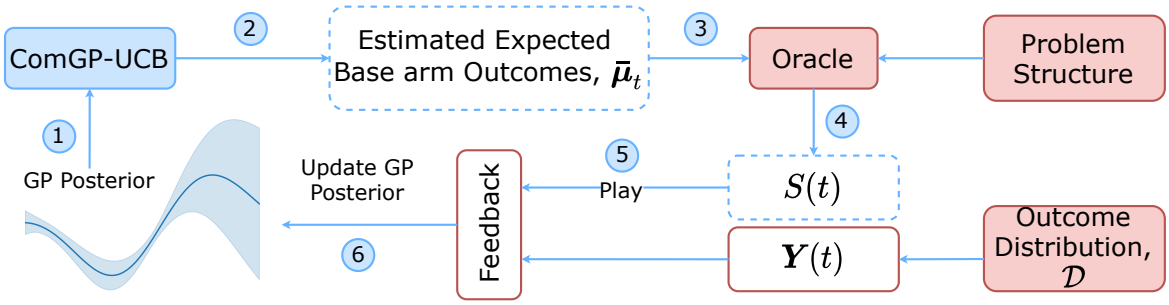

Figure 1: *Algorithm Flow.* At each round t, the environment draws a random outcome vector $\boldsymbol{Y}(t)$ from the true outcome distribution $\mathcal{D}$. ComGP-UCB first forms an optimistic estimate for the expected base arm outcome vector, $\bar{\boldsymbol{\mu}}_t$, using the posterior GP distribution and feeds it to the oracle. The oracle knows the problem structure and it returns an optimal super-arm $S(t)$ for the given parameter set, $\bar{\boldsymbol{\mu}}_t$. Then, the learner plays the super-arm $S(t)$ and observes a feedback, $Q(t)$, which depends on $\boldsymbol{Y}(t)$. Finally, the learner updates the GP distribution based on the feedback it observes. Blue boxes are the variables that are created and updated by the learner, whereas the red boxes are external and environmental entities.

regret bound for this case depends on a new term called *probabilistic triggering modulated (pseudo-) information gain.*

Definition 1. *(Probabilistic Triggering Modulated Pseudo-Information Gain)*

$$\gamma^{PTA}_{T,\boldsymbol{\mu}} := \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in\tilde{S}(t)}(p_i^{S(t)})^2\log(1+\sigma^{-2}\hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i)). \tag{4}$$

Probabilistic triggering modulated pseudo-information gain is very similar to the standard information gain term (Srinivas et al., 2010). In the next lemma, we consider the CMAB framework with no probabilistically triggered arms, in which the super-arm S is constructed via a greedy approach. That is, within round t, the learner selects a single base arm, observes its outcome, and updates the algorithm's parameters before choosing the next base arm. We consider such a case to provide an analogy between the standard information gain term and *probabilistic triggering modulated (pseudo-) information gain* term. More details can be found in supplementary material.

Lemma 1. *The information gain by the end of round T in a CMAB problem (no probabilistic triggering) where the super-arm S is constructed with a greedy approach can be expressed as,*

$$I(\boldsymbol{y}_{[T]};\boldsymbol{f}_{[T]}) = \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S'(t)}\log(1+\sigma^{-2}\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)), \tag{5}$$

where $\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)$ is the conditional variance at $\boldsymbol{x}_i\in\mathcal{X}$ conditioned on all the selected base arms before ith arm at round t.

However, there are two crucial differences between the two terms. The first difference is that we sum over the triggering set while calculating the pseudo-information gain, $\tilde{S}(t)$. Even though a base arm is not triggered, it still contributes to the sum, a situation arising from the probabilistic triggering part of the considered problem. Another difference is that ComGP-UCB executes batch updates on the mean vector and covariance matrix; that is, they are not updated after every single base arm is selected and its outcome is observed but after a round is completed. Despite the differences, pseudo-information gain essentially characterizes a piece of similar information with the standard information gain term (Srinivas et al., 2010). Although we do not provide explicit theoretical bounds on the pseudo-information gain term, we empirically demonstrate how it grows over time in Section 5. Now we are ready to state the next theorem.

Theorem 2. *Given $\delta\in(0,1)$, $C := \frac{8\tilde{B}^2}{\log(1+\sigma^{-2})}$, where $\tilde{B}$ is the TPM Lipschitz constant, and under Assumptions 1, 2, and 4, the cumulative regret of ComGP-UCB after round T is upper bounded with at least $1-\delta$ probability as follows,*

$$\mathbb{P}\{Reg_{\boldsymbol{\mu}}(T)\le\sqrt{CmT\beta_{mT}\gamma^{PTA}_{T,\boldsymbol{\mu}}}\}\ge 1-\delta.$$

The bound in Theorem 2 matches Theorem 1 in Srinivas et al. (2010) in terms of its dependence on time. However, in Theorem 2, we have an extra $\sqrt{m}$ term due to the combinatorial nature of the problem, and we have the pseudo-information gain term instead of the standard information gain term. Note that, one can also use the vanilla GP-UCB algorithm to solve a CMAB problem by treating each super arm as a single base arm, where the context vector of each super-arm is defined as the concatenation of its base arms' context vectors. In this scenario, the $\sqrt{m}$ term in the

regret bound will disappear. However, this approach has one immediate drawback: the information related to underlying base arms shared between different super arms will be discarded. Another drawback is that the number of possible super arms increases combinatorially with respect to the number of base arms and the GP-based inference on a high-dimensional space might degrade the performance.

4.2 Preliminaries for the Proofs

Let us define the following events,

$$H_i(t) \coloneqq \{i \in \tilde{S}(t), N_i(t) \leq (1-\rho)p_i M_i(t)\}$$
$$H(t) \coloneqq \{\exists i \in [m] : H_i(t)\}$$
$$\mathcal{G} \coloneqq \{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i| \leq \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i), \forall i \in [m], \forall t \in [T]\}$$
$$\mathcal{J} \coloneqq \Big\{\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\} < \alpha\Big\},$$

where $\alpha \in [mT]$ is fixed, $M_i(t) \coloneqq \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in \tilde{S}(\tau)\}$ denotes the number of times base arm i was in the triggering set, and $N_i(t) \coloneqq \sum_{\tau=1}^{t-1}\mathbb{I}\{i \in S'(\tau)\}$ denotes the number of observations made at base arm i, up to round t. Also, let τ_w^i denote the round number where ith base arm is in the triggering set of the selected super arm for the wth time, and $\tau_0^i = 0$.

4.3 Supplemental Lemmas

We have the following three lemmas for which the detailed proofs can be found in the supplementary material.

Lemma 2. *Given $\delta \in (0,1)$, the event $\mathcal{G}$ holds with at least $1-\delta$ probability.*

Lemma 3. *Given $\alpha \in [mT]$, the event $\mathcal{J}$ holds with at least $1 - \frac{m}{\alpha} - \frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$ probability.*

Lemma 4. *Instantaneous variance of the Gaussian process at $\boldsymbol{x}_i \in \mathcal{X}$ can be upper bounded by the noise variance and the number of times base arm i was triggered as, $\sigma^2/N_i(t) \geq \hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i)$.*

4.4 Proof of Theorem 1

For the proof of Theorem 1, we assume that the events $\mathcal{G}$ and $\mathcal{J}$ hold, which means that $\bar{\mu}_t(\boldsymbol{x}_i) \geq \mu_i$ for all $i \in [m]$ and $t \in [T]$, and $\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\} < \alpha$ for a fixed $\alpha \in [mT]$.

$$Reg_{\boldsymbol{\mu}}(T) = \sum_{t=1}^{T} r(S^*, \boldsymbol{\mu}) - r(S(t), \boldsymbol{\mu})$$
$$\leq \sum_{t=1}^{T} r(S^*, \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu}) \quad (6)$$
$$\leq \sum_{t=1}^{T} r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu}) \quad (7)$$
$$\leq \sum_{t=1}^{T} \mathbb{I}\{\neg H(t)\}(r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu})) \quad (8)$$
$$+ \sum_{t=1}^{T} \mathbb{I}\{H(t)\} \times 2mB\sqrt{\beta_{mT}}, \quad (9)$$

where (6) follows from Assumption 2, (7) is due to the definition of the OPT operator, and (9) is obtained by observing that for all $S \in \mathcal{I}$,

$$|r(S, \bar{\boldsymbol{\mu}}_t) - r(S, \boldsymbol{\mu})|$$
$$\leq B\sum_{i\in\tilde{S}} |\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)| \quad (10)$$
$$\leq B\sum_{i\in\tilde{S}} \big(|\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i)| + |\sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)|\big)$$
$$\leq 2B\sum_{i\in\tilde{S}} \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \quad (11)$$
$$\leq 2mB\sqrt{\beta_{mT}}, \quad (12)$$

where (10) follows from Assumption 3, (11) holds since $\mathcal{G}$ holds, and (12) holds since $mT \geq N(t)$ for all $t \in [T]$, and $\hat{\sigma}^2_{N(t)}(x_i) \leq 1$ since we assume bounded variance.

Bounding (8). When $\mathcal{G}$ and $\neg H(t)$ hold, we have,

$$|r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu})|$$
$$\leq 2B \sum_{i\in\tilde{S}(t)} \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \quad (13)$$
$$\leq 2B \sum_{i\in\tilde{S}(t)} \sqrt{\beta_{N(t)}}\frac{\sigma}{\sqrt{N_i(t)}} \quad (14)$$
$$\leq 2B \sum_{i\in\tilde{S}(t)} \sqrt{\beta_{N(t)}}\frac{\sigma}{\sqrt{(1-\rho)p_i M_i(t)}}, \quad (15)$$

where (13) follows from (11), (14) follows from Lemma 4, and (15) holds since $\neg H(t)$ holds. Then,

$$(8) \leq \sum_{t=1}^{T}\mathbb{I}\{\neg H(t)\}2B \sum_{i\in\tilde{S}(t)} \sqrt{\beta_{N(t)}}\frac{\sigma}{\sqrt{(1-\rho)p_i M_i(t)}}$$
$$\leq \sqrt{\beta_{mT}}\sum_{t=1}^{T}\mathbb{I}\{\neg H(t)\}2B \sum_{i\in\tilde{S}(t)} \frac{\sigma}{\sqrt{(1-\rho)p_i M_i(t)}}$$
$$\leq \sqrt{\beta_{mT}}\sum_{i=1}^{m}\sum_{w=0}^{T}\sum_{t=\tau_w^i+1}^{\tau_{w+1}^i} \mathbb{I}\{i \in \tilde{S}(t), \neg H(t)\} \times 2B\frac{\sigma}{\sqrt{(1-\rho)p_i M_i(t)}}$$
$$\leq \sqrt{\beta_{mT}}\sum_{i=1}^{m}\sum_{w=0}^{T} 2B\frac{\sigma}{\sqrt{(1-\rho)p_i M_i(\tau_{w+1}^i)}}$$

$$\leq \sqrt{\beta_{mT}} \sum_{i=1}^{m} \sum_{w=1}^{T} 2B \frac{\sigma}{\sqrt{(1-\rho)p^* w}} \tag{16}$$

$$\leq 4mB \sqrt{\frac{T\beta_{mT}\sigma^2}{(1-\rho)p^*}}, \tag{17}$$

where (16) holds since $M_i(\tau_1^i) = N_i(\tau_1^i) = 0$ for all i, meaning $\neg H(\tau_1^i)$ does not hold. Finally, (17) follows from the fact $\sum_{n=1}^{N} \sqrt{1/n} \leq 2\sqrt{N}$.

Bounding (9).

$$(9) = 2mB\sqrt{\beta_{mT}} \sum_{t=1}^{T} \mathbb{I}\{H(t)\}$$

$$\leq 2mB\sqrt{\beta_{mT}} \sum_{i=1}^{m} \sum_{t=1}^{T} \mathbb{I}\{H_i(t)\}$$

$$\leq 2m\alpha B\sqrt{\beta_{mT}}, \tag{18}$$

where (18) holds since $\mathcal{J}$ holds. Finally, when the events $\mathcal{G}$ and $\mathcal{J}$ hold, we can bound the cumulative regret after T rounds as follows,

$$Reg_{\boldsymbol{\mu}}(T) \leq (8) + (9)$$

$$\leq 4mB \sqrt{\frac{T\beta_{mT}\sigma^2}{(1-\rho)p^*}} + 2m\alpha B\sqrt{\beta_{mT}}.$$

Theorem 1 follows by combining Lemmas 2 and 3 with the fact that $\mathbb{P}\{\mathcal{G} \cap \mathcal{J}\} = 1 - \mathbb{P}\{\neg\mathcal{G} \cup \neg\mathcal{J}\}$, where $1 - \mathbb{P}\{\neg\mathcal{G} \cup \neg\mathcal{J}\} \geq 1 - \mathbb{P}\{\neg\mathcal{G}\} - \mathbb{P}\{\neg\mathcal{J}\}$, by union bound. Finally, $1 - \mathbb{P}\{\neg\mathcal{G}\} - \mathbb{P}\{\neg\mathcal{J}\} = \mathbb{P}\{\mathcal{G}\} + \mathbb{P}\{\mathcal{J}\} - 1$, and $\mathbb{P}\{\mathcal{G}\} + \mathbb{P}\{\mathcal{J}\} - 1 \geq 1 - \delta - \frac{m}{\alpha} - \frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$.

4.5 Proof of Theorem 2

For the proof of Theorem 2, we assume that the event $\mathcal{G}$ holds.

$$Reg_{\boldsymbol{\mu}}(T) \leq \sum_{t=1}^{T} r(S(t), \bar{\boldsymbol{\mu}}_t) - r(S(t), \boldsymbol{\mu}) \tag{19}$$

$$\leq 2\tilde{B} \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_i^{S(t)} \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \tag{20}$$

$$\leq 2\tilde{B}\sqrt{\beta_{mT}} \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_i^{S(t)} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i), \tag{21}$$

where (19) follows from (7), (21) holds since $mT \geq N(t)$ for all $t \in [T]$, and (20) is obtained by observing that for all $S \in \mathcal{I}$,

$$|r(S, \bar{\boldsymbol{\mu}}_t) - r(S, \boldsymbol{\mu})|$$

$$\leq \tilde{B} \sum_{i \in \tilde{S}} p_i^S |\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i)| \tag{22}$$

$$\leq \tilde{B} \sum_{i \in \tilde{S}} p_i^S \left(|\mu_i - \hat{\mu}_{N(t)}(\boldsymbol{x}_i)| + \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \right)$$

$$\leq 2\tilde{B} \sum_{i \in \tilde{S}} p_i^S \sqrt{\beta_{N(t)}} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i), \tag{23}$$

where (22) follows from Assumption 4, and (23) holds since $\mathcal{G}$ holds. Let us define $p_{i,S(t)} = p_i^{S(t)}$ to keep the notation uncluttered. By Cauchy-Schwartz inequality, we can write,

$$Reg_{\boldsymbol{\mu}}^2(T) \leq 4\tilde{B}^2 T\beta_{mT} \sum_{t=1}^{T} \Big(\sum_{i \in \tilde{S}(t)} p_{i,S(t)} \hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \Big)^2$$

$$\leq 4\tilde{B}^2 mT\beta_{mT} \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_{i,S(t)}^2 \hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i) \tag{24}$$

$$= 4\tilde{B}^2 mT\beta_{mT}\sigma^2 \times \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_{i,S(t)}^2 \sigma^{-2} \hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i)$$

$$\leq 8\tilde{B}^2 mT\beta_{mT} \times \frac{1}{2} \sum_{t=1}^{T} \sum_{i \in \tilde{S}(t)} p_{i,S(t)}^2 \frac{\log(1 + \sigma^{-2}\hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i))}{\log(1+\sigma^{-2})} \tag{25}$$

$$= CmT\beta_{mT}\gamma_{T,\boldsymbol{\mu}}^{PTA}, \tag{26}$$

where $C := \frac{8\tilde{B}^2}{log(1+\sigma^{-2})}$, is a constant, and $\gamma_{T,\boldsymbol{\mu}}^{PTA}$ is *pseudo-information gain* in Definition 1. (24) follows applying Cauchy-Schwarz inequality again, (25) follows from the fact that $s^2 \leq \frac{\sigma^{-2} log(1+s^2)}{log(1+\sigma^{-2})}$ for $s \in [0, \sigma^{-2}]$, and $\sigma^{-2}\hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i) \leq \sigma^{-2}$, since $\hat{\sigma}_{N(t)}^2(\boldsymbol{x}_i) \leq 1$.

5 NUMERICAL RESULTS

5.1 Cascading Bandit

We compare the performance of ComGP-UCB against other state-of-the-art algorithms in a disjunctive cascading bandit problem (Kveton et al., 2015b), which can be treated as a CMAB-PTA instance (Wang et al., 2017). We consider an item list recommendation scenario, where a search engine tries to recommend the most attractive list of webpages to its users. The users come one at a time, and the engine selects K pages from a set of R pages for each user. Each K-sized super-arm (i.e., S) is modeled as an immediately triggered virtual arm. The base arms in virtual arm S that can be triggered constitute $\tilde{S}$. The oracle (OPT) in this experiment returns an ordered list of K best webpages from R webpages based on their estimated click probabilities at each round t. (i.e., $\bar{\boldsymbol{\mu}}_t$). The user examines the pages in the order they were recommended and clicks on the first webpage she finds interesting.

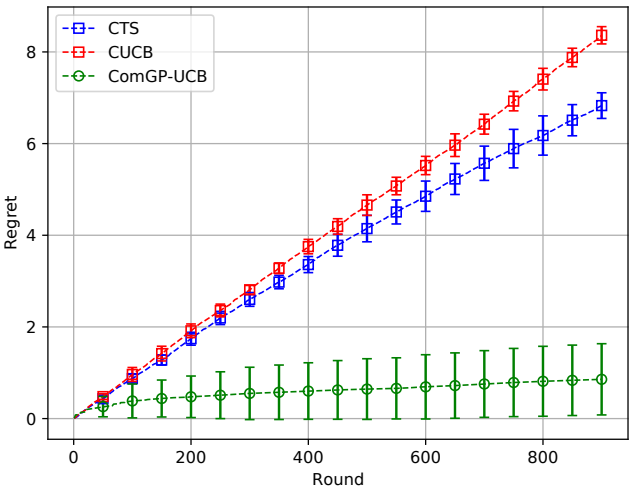

Figure 2: Regrets for the disjunctive cascading bandit problem (±2 standard deviation), when there is high correlation between expected base arm outcomes.

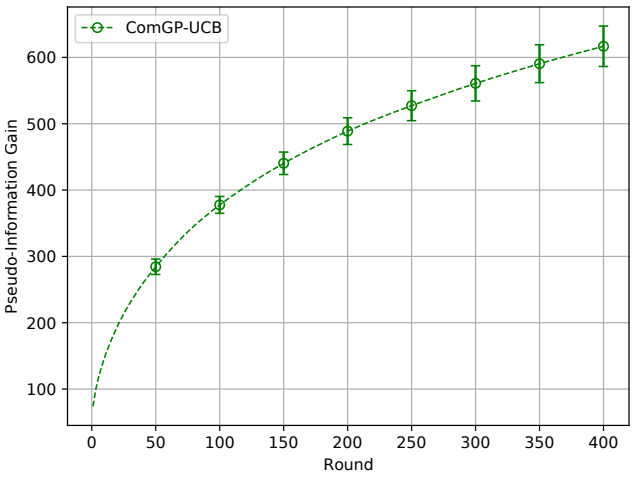

Figure 3: Pseudo-information gain term, $\gamma^{PTA}_{T,\boldsymbol{\mu}}$, when ComGP-UCB is run in the disjunctive cascading bandit problem (±2 standard deviation), when there is high correlation between expected base arm outcomes.

If a user does not find any of the pages attractive, she does not click on any page. The search engine aims to maximize the total number of clicks.

We treat each webpage $j \in [R]$ as a base arm. The user finds the page j interesting with probability p_j. The super arms are lists of size K, consisting of the webpages recommended to the user. Let $S(k)$ denote the kth page recommended to the user, given a super arm S. Then, we can express the triggering probabilities for each base arm in S as,

$$p^S_{(j)} = \begin{cases} 1 & \text{if } j = S(1) \\ \prod_{k'=1}^{k-1}(1 - p_{S(k')}) & \text{if } \exists k \neq 1 : j = S(k) \\ 0 & \text{otherwise} \end{cases}$$

We observe feedback for the first recommendations (pages) immediately and observe feedback for the other ones only if the user does not click on the previous recommendations. Then, the expected reward of playing super arm S can be written as,

$$r(S, \boldsymbol{p}) = \left(1 - \prod_{k=1}^{K}(1 - p_{S(k)})\right)$$

for which Assumptions 3 and 4 hold when $B = 1$ and $\tilde{B} = 1$. We consider a scenario where the engine recommends 5 pages among 900 pages to a user, that is, $R = 900$ and $K = 5$. We assume that each webpage has a 2-dimensional context vector, $\boldsymbol{x}$, and generate p_j's by first sampling from a Gaussian process with squared-exponential kernel[2] on a 2D grid, and then passing the output through a sigmoid function (van der Wilk et al., 2020). We run ComGP-UCB, CUCB in Chen et al. (2016), and CTS in Hüyük et al. (2019) for 900 rounds and average their regrets over 10 different runs. Note that CTS and CUCB do not use the side information (context) related to webpages to exploit the dependencies between their click probabilities. In Figure 2, ComGP-UCB is clearly superior to the other algorithms. That is because the other algorithms assume independence between expected base arm outcomes, whereas ComGP-UCB utilizes the information from every base arm observation to update its knowledge about the other base arms. On the other hand, we observe in Figure 4 that when we sample p_js randomly between $[0, 1]$, ComGP-UCB is not significantly superior, as expected. In Figure 3, we have the pseudo-information gain term, $\gamma^{PTA}_{T,\boldsymbol{\mu}}$, which increases sublinearly in time.

[2]We set variance parameter to 2, lengthscale parameter to 0.8, and likelihood variance to 0.01.

5.2 Sparse GPs and Kernel Parameters

One disadvantage of the GPs is that they are computationally expensive. ComGP-UCB has a computational complexity of $O(n^3)$, where n denotes the total number of base arm observations. We can always use sparse GPs for inference to reduce the computational complexity to $O(m^2 n)$, where m is a chosen constant (number of inducing points) (Titsias, 2009). For the sake of demonstration, we consider a synthetic problem where the reward of a super-arm S is simply the sum of its base arms' individual rewards. Formally, $R(S(t), \boldsymbol{Y}(t)) = \sum_{i \in S(t)} \bar{r}^t_i$, where $\bar{r}^t_i$ denotes the outcome of base arm $i \in [m]$ at round t. We show in Figure 5 that sparse GP still significantly outperforms CTS and CUCB and it has a competitive performance

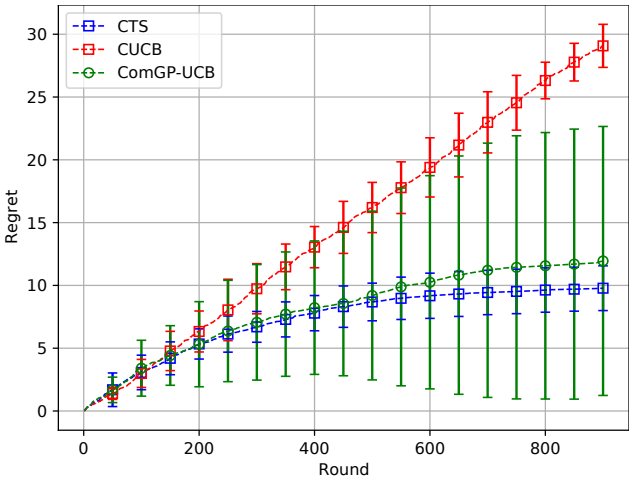

Figure 4: Regrets for the disjunctive cascading bandit problem (±2 standard deviation), when expected base arm outcomes are sampled randomly.

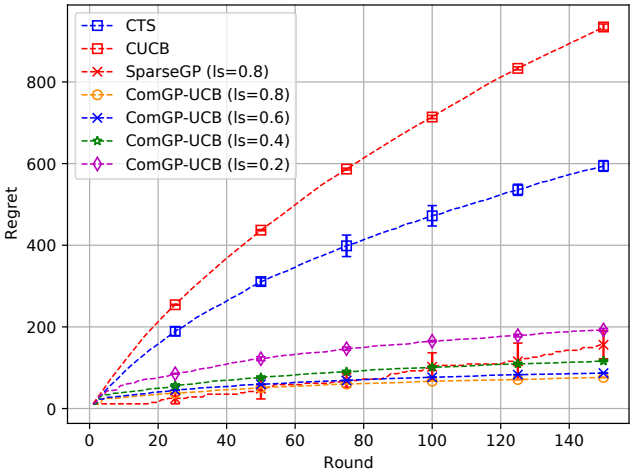

Figure 5: Regrets for synthetic linear reward problem (±2 standard deviation), when the true kernel (squared-exponential kernel) lengthscale (ls) is set to 0.8.

against ComGP-UCB. Moreover, we assume that the kernel hyper-parameters are known a priori. In reality, these parameters can be estimated from a training dataset or can be tuned by using domain knowledge. We show in Figure 5 that ComGP-UCB manages to outperform its competitors even when there is a hyper-parameter mismatch; that is, the GP does not know the true kernel hyper-parameters a priori.

6 CONCLUSION

We analyzed the regret of the ComGP-UCB algorithm in a CMAB-PTA setting and derived high probability upper bounds on its regret under (TPM) Lipschitz continuity and monotonicity assumptions on the expected reward function. One of our bounds includes the $1/p^*$ term that is unavoidable in general, and the other one includes a pseudo-information gain term for which we have provided simulations regarding its behavior in time. Interesting future work includes a more detailed investigation of the pseudo-information gain and providing explicit upper bounds. Overall, we showed that when the expected base arm outcomes are not independent but a sample from a GP, ComGP-UCB algorithm significantly outperforms the state-of-the-art benchmarks CTS and CUCB.

Acknowledgements

This work was supported in part by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 215E342 and in part by the BAGEP Award of the Science Academy.

References

D. A. Berry and B. Fristedt, *Bandit problems: sequential allocation of experiments (Monographs on statistics and applied probability)*., 1985

R. S. Sutton, A. G. Barto, et al., *Introduction to reinforcement learning*, vol. 135, 1998.

H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 55, pp. 527–535, 1952.

T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, pp. 4–22, 1985.

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.

S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," *Proceedings of the 25th Annual Conference on Learning Theory*, pp. 39.1–39.26, 2012.

D. Russo and B. V. Roy, "Learning to optimize via posterior sampling," *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

R. Agrawal, "Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables:

multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1466–1478, 2012.

B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 535–543, 2015.

W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: general framework and applications," *Proceedings of the 30th International Conference on Machine Learning*, pp. 151–159, 2013.

A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," *Proceedings of the 31st International Conference on Machine Learning*, pp. 100–108, 2014.

A. Hüyük and C. Tekin, "Thompson sampling for combinatorial network optimization in unknown environments" *IEEE/ACM Transactions on Networking*, vol. 28, no. 6, pp. 2836–2849, 2020.

W. Chen, Y. Wang, Y. Yuan, and Q. Wang, "Combinatorial multi-armed bandit and its extension to probabilistically triggered arms," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1746–1778, 2016.

Q. Wang and W. Chen, "Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications," *Advances in Neural Information Processing Systems*, pp. 1161–1171, 2017.

S. Wang and W. Chen, "Thompson sampling for combinatorial semi-bandits," *Proceedings of the 35th International Conference on Machine Learning*, pp. 5114–5122, 2018.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *Proceedings of the 27th International Conference on Machine Learning*, pp. 1015–1022, 2010.

C. E. Rasmussen, "Gaussian processes in machine learning," *Advanced Lectures on Machine Learning*, 2004.

B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan, "Cascading bandits: learning to rank in the cascade model," *Proceedings of the 32nd International Conference on Machine Learning*, pp. 767–776, 2015.

A. Hüyük and C. Tekin, "Analysis of Thompson sampling for combinatorial multi-armed bandit with probabilistically triggered arms," *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1322–1330, 2019.

W. C. Cheung, V. Tan, and Z. Zhong, "A Thompson sampling algorithm for cascading bandits," *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 438–447, 2019.

M. van der Wilk, V. Dutordoir, S. John, A. Artemev, V. Adam, and J. Hensman, "A framework for interdomain and multioutput Gaussian processes," *arXiv:2003.01115*, 2020.

M. Titsias, "Variational learning of inducing variables in sparse Gaussian processes," *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 567–574, 2009.