A SUPPLEMENTAL DOCUMENT

A.1 PROOFS OF LEMMAS

A.1.1 Preliminaries

Let us remember the definitions of the events,

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

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 $\tau_0^i = 0$.

Also, let us remind the definition of the exploration coefficient, $\beta_n = 2\log(\frac{mT\pi_n}{\delta})$, where $\sum_{n=0}^{\infty} \frac{1}{\pi_n} = 1$ (e.g., $\pi_n = \pi^2 n^2/6$). Note that β_n is strictly increasing in n.

Fact 1. *(Multiplicative Chernoff bound (Chen et al., 2016)) Let $X_1, \ldots, X_n$ be Bernoulli random variables taking values in $\{0,1\}$ such that $\mathbb{E}[X_t|X_1, \ldots, X_{t-1}] \geq \mu$ for all $t \leq n$, and $Y = X_1 + \cdots + X_n$. Then, for all $\delta \in (0,1)$,*

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

A.1.2 Proof of Lemma 1

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,*

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

where $N_{(t,i)}$ denotes the number of observations up to (not including) ith observation in round t, and $\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)$ denotes the conditional variance at x_i, conditioned on all the selected base-arms before ith arm at round t.

Proof. The main goal of this lemma is to provide an analogy between the standard information gain, and the pseudo-information gain term that appears in our regret analysis. For the proof of this lemma, we consider a scenario the where base arms are selected with a greedy approach, that is, after a base arm is chosen, the statistics of the learner's policy are updated before choosing the next base arm in the same round. As aforementioned, this lemma serves to provide an analogy for our pseudo-information gain term. Even though this assumption may not hold in general, there are several cases it holds. For instance, in the disjunctive form of cascading bandit problem, where a search engine recommends webpages to its users, we can recommend webpages to a user one at a time and update the GP statistics before our next recommendation. The proof this lemma is inspired by the proof of Lemma 5.3 in Srinivas et al. (2010). Let $\boldsymbol{x}_t^k$ denote the kth arm in round t for which the learner observes a feedback, and y_t^k denote the observation at $\boldsymbol{x}_t^k$. Furthermore, let $\boldsymbol{y}_t^k$ be the vector of observations up to (not including) kth observation in round t, and n_t denote the total number of observations in round t. Also, let N denote the total number of observations by the end of round T. Conditioned on $\boldsymbol{y}_t^k$, $\boldsymbol{x}_1^1, \boldsymbol{x}_1^2, \ldots, \boldsymbol{x}_t^k$ are deterministic, and the conditional variance $\hat{\sigma}^2_{N(t,k)}(\boldsymbol{x}_t^k)$ does not depend on $\boldsymbol{y}_t^k$. Then, we have,

$$\begin{aligned} I(\boldsymbol{y}_T; \boldsymbol{f}_T) &= H(\boldsymbol{y}_T) - \frac{1}{2}\log|2\pi e\sigma^2 \boldsymbol{I}| \\ &= H(\boldsymbol{y}_T) - \frac{N}{2}\log(2\pi e\sigma^2) \end{aligned} \tag{2}$$

where $\boldsymbol{y}_T$ denotes the vector of observations by the end of round T. We can rewrite $H(\boldsymbol{y}_T)$ as,

$$\begin{aligned}
H(\boldsymbol{y}_T) &= H(\boldsymbol{y}_T^{n_T}) + H(y_T^{n_T}|\boldsymbol{y}_T^{n_T}) \\
&= H(\boldsymbol{y}_T^{n_T-1}) + H(y_T^{n_T-1}|\boldsymbol{y}_T^{n_T-1}) + H(y_T^{n_T}|\boldsymbol{y}_T^{n_T}) \\
&= H(\boldsymbol{y}_1^2) + H(y_1^2|\boldsymbol{y}_1^2) + \ldots + H(y_T^{n_T-1}|\boldsymbol{y}_T^{n_T-1}) + H(y_T^{n_T}|\boldsymbol{y}_T^{n_T}) \\
&= \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S'(t)} \log(2\pi e\sigma^2(1+\sigma^{-2}\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)) \\
&= \frac{N}{2}\log(2\pi e\sigma^2) + \frac{1}{2}\sum_{t=1}^{T}\sum_{i\in S'(t)} \log(1+\sigma^{-2}\hat{\sigma}^2_{N(t,i)}(\boldsymbol{x}_i)) \qquad (3)
\end{aligned}$$

Finally, by the end of round T, given a set of base-arm observations $\boldsymbol{y}_T$, the information gain expression in (1) follows from combining (2) and (3). □

A.1.3 Proof of Lemma 2

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

Proof.

Fact 2. *Given $r \sim \mathcal{N}(0,1)$, $\mathbb{P}\{|r| > c\} \leq e^{-c^2/2}$, by tail inequality.*

Recall that the information available to the learner to guide its actions in round $t+1$ is $\mathcal{F}_t \coloneqq \{(S(\tau), Q(\tau)) : \tau \in [t]\}$. After making a set of observations, we can update the posterior mean and variance at $\boldsymbol{x} \in \mathcal{X}$ as follows,

$$\begin{aligned}
k_{N(t)}(\boldsymbol{x}, \boldsymbol{x}') &= k(\boldsymbol{x},\ \boldsymbol{x}') \\
&\quad - \boldsymbol{k}_{N(t)}(\boldsymbol{x})^T(\boldsymbol{K}_{N(t)} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{k}_{N(t)}(\boldsymbol{x}') \qquad (4) \\
\hat{\sigma}^2_{N(t)}(\boldsymbol{x}) &= k_{N(t)}(\boldsymbol{x},\ \boldsymbol{x}) \qquad (5) \\
\hat{\mu}_{N(t)}(\boldsymbol{x}) &= \boldsymbol{k}_{N(t)}(\boldsymbol{x})^T(\boldsymbol{K}_{N(t)} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{Y}_t, \qquad (6)
\end{aligned}$$

where $\boldsymbol{Y}_t \coloneqq [\boldsymbol{Y}^T(1), \ldots, \boldsymbol{Y}^T(t-1)]^T$ denotes the vector of observations made until round t, and $N(t) \coloneqq |\boldsymbol{Y}_t|$ 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. $\boldsymbol{k}_{N(t)}(\boldsymbol{x}) \coloneqq [k(\boldsymbol{x}^1,\ \boldsymbol{x}), \ldots, k(\boldsymbol{x}^{N(t)},\ \boldsymbol{x})]^T$ denotes the vector of covariances between $\boldsymbol{x} \in \mathcal{X}$, and past observations $[\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{N(t)}]$, 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.

At round t, we have $\mu_i \mid \mathcal{F}_t \sim \mathcal{N}(\hat{\mu}_{N(t)}(\boldsymbol{x}_i), \hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i))$. By using Fact 2, we can write,

$$\begin{aligned}
\mathbb{P}\big\{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i| > \sqrt{\beta_n}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \mid \mathcal{F}_t\big\} &= \mathbb{P}\big\{\frac{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i|}{\hat{\sigma}_{N(t)}(\boldsymbol{x}_i)} > \sqrt{\beta_n} \mid \mathcal{F}_t\big\} \\
&\leq e^{-\beta_n/2}
\end{aligned}$$

for a fixed $i \in [m]$ and n. Calculating the union bound over all m base arms, time horizon T, and number of

arms that can be chosen up to round t, $N(t)$, we observe

$$\begin{aligned}
&\bigcup_{i,t,n} \mathbb{P}\big\{|\hat{\mu}_{N(t)}(\boldsymbol{x}_i) - \mu_i| > \sqrt{\beta_{N(t)}}\hat{\sigma}_{N(t)}(\boldsymbol{x}_i) \cap N(t) = n \mid \mathcal{F}_t\big\} \\
&\qquad \leq \sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{n=0}^{m(t-1)} e^{-\beta_n/2} P\{N(t) = n \mid \mathcal{F}_t\} \\
&\qquad \leq \sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{n=0}^{\infty} \frac{\delta}{mT\pi_n} \qquad (7) \\
&\qquad \leq \frac{\delta}{mT}\sum_{i=1}^{m}\sum_{t=1}^{T}\sum_{n=0}^{\infty} \frac{1}{\pi_n} \\
&\qquad \leq \delta \sum_{n=0}^{\infty} \frac{1}{\pi_n} \\
&\qquad \leq \delta \qquad (8)
\end{aligned}$$

where (7) follows from the definition of β_n, and (8) follows from fact that $\sum_{n=0}^{\infty} 1/\pi_n = 1$. Finally, by definition of the event $\mathcal{G}$, and (8), we can observe that $\mathbb{P}\big\{\mathcal{G}\big\} \geq 1 - \delta$. □

A.1.4 Proof of Lemma 3

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.*

Proof.

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\{i \in \tilde{S}(t), N_i(t) \leq (1-\rho)p_i M_i(t)\}\middle|\boldsymbol{\mu}\right] \\
&\quad \leq \mathbb{E}\left[\sum_{w=0}^{T}\sum_{t=\tau_w^i+1}^{\tau_{w+1}^i} \mathbb{I}\{i \in \tilde{S}(t),\ N_i(t) \leq (1-\rho)p_i M_i(t)\}\middle|\boldsymbol{\mu}\right] \\
&\quad \leq \mathbb{E}\left[\sum_{w=0}^{T} \mathbb{I}\{N_i(\tau_{w+1}^i) \leq (1-\rho)p_i M_i(\tau_{w+1}^i)\}\middle|\boldsymbol{\mu}\right] \\
&\quad \leq 1 + \sum_{w=1}^{T} \mathbb{P}\{N_i(\tau_{w+1}^i) \leq (1-\rho)p_i M_i(\tau_{w+1}^i)|\boldsymbol{\mu}\} \\
&\quad \leq 1 + \sum_{w=1}^{T} e^{-\frac{\rho^2 p^* w}{2}} \qquad (9) \\
&\quad \leq 1 + \frac{2}{\rho^2 p^*}
\end{aligned}$$

where (9) is due to Fact 1. We can then write,

$$\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{I}\{H_i(t)\} \mid \boldsymbol{\mu}\Big] \leq 1 + \frac{2}{\rho^2 p^*}\ .$$

Then, we have

$$\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T} \mathbb{I}\{H_i(t)\} \mid \boldsymbol{\mu}\Big] \leq m + \frac{2m}{\rho^2 p^*}\ . \qquad (10)$$

We can now use (10) and tower rule to observe

$$\mathbb{E}_{\boldsymbol{\mu}}\big[\mathbb{E}[\sum_{i=1}^{m}\sum_{t=1}^{T} \mathbb{I}\{H_i(t)\} \mid \boldsymbol{\mu}]\big] \leq m + \frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]$$

which yields,

$$\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\Big] \leq m + \frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}] \ . \tag{11}$$

Since $\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}$ is non-negative, we can use Markov's inequality for a fixed $\alpha \in [mT]$ and write

$$\mathbb{P}\big\{\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\} \geq \alpha\big\} \leq \frac{\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\Big]}{\alpha} \ . \tag{12}$$

Finally, we combine the definition of event $\mathcal{J}$, (11), and (12) to observe

$$\begin{aligned}
\mathbb{P}\big\{\mathcal{J}\big\} &= \mathbb{P}\big\{\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\} < \alpha\big\} \\
&\geq 1 - \frac{\mathbb{E}\Big[\sum_{i=1}^{m}\sum_{t=1}^{T}\mathbb{I}\{H_i(t)\}\Big]}{\alpha} \\
&\geq 1 - \frac{m + \frac{2m}{\rho^2}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}]}{\alpha} \\
&\geq 1 - \frac{m}{\alpha} - \frac{2m}{\rho^2\alpha}\mathbb{E}_{\boldsymbol{\mu}}[\frac{1}{p^*}] \ .
\end{aligned}$$

□

A.1.5 Proof of Lemma 4

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, $\sigma^2/N_i(t) \geq \hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i)$.*

Proof. Given $A \subseteq B$, $H(\mu|A) \geq H(\mu|B)$, since conditioning on more observations will reduce entropy. Let $\boldsymbol{Y}_t^i$ denote the observations made at base arm i up to round t. Remember that $\boldsymbol{Y}_t$ denotes all the observations up to round t. Then, we have $\boldsymbol{Y}_t^i \subseteq \boldsymbol{Y}_t$, which implies $H(\mu|\boldsymbol{Y}_t^i) \geq H(\mu|\boldsymbol{Y}_t)$. The entropy of a Gaussian random variable is $H(\mathcal{N}(\mu,\sigma^2)) = \frac{1}{2}\log(2\pi e\sigma^2)$. We have $H(\mu|\boldsymbol{Y}_t) = \frac{1}{2}\log(2\pi e\hat{\sigma}^2_{N(t)})$, and $H(\mu|\boldsymbol{Y}_t^i) = \frac{1}{2}\log(2\pi e(N_i(t)/\sigma^2+\sigma_i^{-2})^{-1}))$. Then we can write,

$$\frac{1}{2}\log(\frac{2\pi e}{N_i(t)/\sigma^2 + \sigma_i^{-2}}) \geq \frac{1}{2}\log(2\pi e\hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i)) \tag{13}$$

where $\sigma_i^2 = k(\boldsymbol{x}_i,\boldsymbol{x}_i)$ is the prior variance of GP at $\boldsymbol{x}_i$. Since $k(\boldsymbol{x}_i,\boldsymbol{x}_i) \geq 0$, the following is immediate from (13),

$$\frac{\sigma^2}{N_i(t)} \geq \hat{\sigma}^2_{N(t)}(\boldsymbol{x}_i) \ .$$

□

A.2 ADDITIONAL NUMERICAL RESULTS

A.2.1 Disjunctive Form of Cascading Bandit Problem

This section provides additional numerical results on the disjunctive form of cascading bandit problem, where a search engine tries to maximize the number of clicks on recommended webpages to its users.

In Figure 1, we have a slightly different setting in which there are multiple users (8) being served simultaneously. Each user-page pair is treated as a base-arm, and each base arm has a two-dimensional context vector. One dimension of the context vector represents the webpage, and the other represents the user. We assume that the expected base-arm outcomes are obtained by passing the output of a GP through sigmoid. We see that when there is a high correlation between base-arms (length-scale = 0.8), ComGP-UCB significantly outperforms other

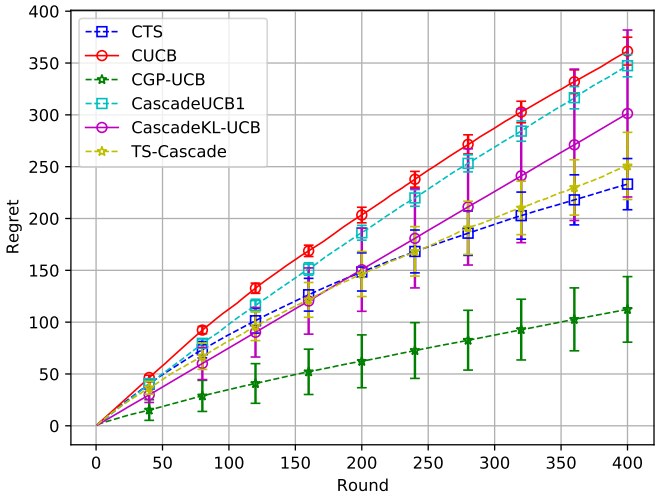

Figure 1: Regrets for the disjunctive cascading bandit problem, when there is high correlation between expected base-arm outcomes. (L=32, R=8, K=4. Variance parameter is set to 1, and length-scale parameter is set to 0.8 for Squared-Exponential Kernel. Likelihood variance is set to 0.01).

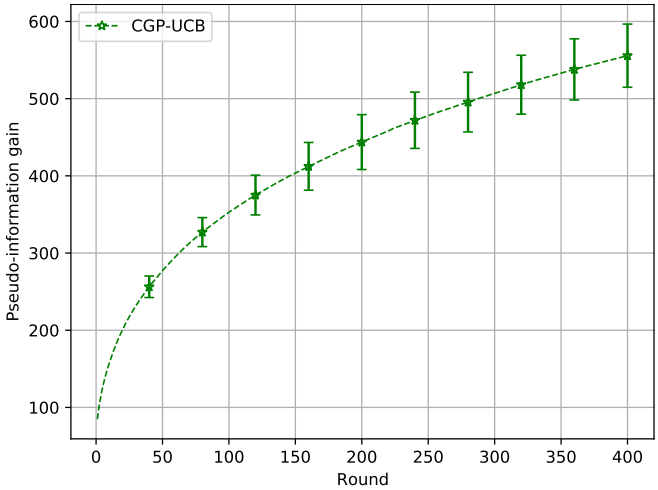

Figure 2: Pseudo-information gain when ComGP-UCB is run (same experiment with Figure 1).

algorithms. In Figure 2, the pseudo-information gain related to ComGP-UCB algorithm in this setting can be observed, which increases sublinearly in time.

Figure 3 shows the regrets of algorithms when the expected base-arm outcomes are sampled randomly. We have two things to note here. First, we observe that the performance ComGP-UCB algorithm is not as good before compared to other algorithms. That is because there is no underlying dependence between different base-arms, which would give ComGP-UCB an advantage over other algorithms. Second, notice two different ComGP-UCB algorithms in Figure 3. The only difference between the two algorithms is that they use a different length-scale (ls) parameter for the squared-exponential kernel. ComGP-UCB1 uses $ls = 0.8$, and ComGP-UCB2 uses $ls = 0.2$. Since a bigger length-scale parameter implies a stronger dependence, ComGP-UCB1 assumes that similar base-arms are highly correlated, whereas ComGP-UCB2 assumes a weaker dependence. Since the expected base-arm outcomes are sampled randomly, ComGP-UCB2 performs better compared to ComGP-UCB1. This observation reveals the importance of working with the correct type of kernel and kernel hyper-parameters for better performance.

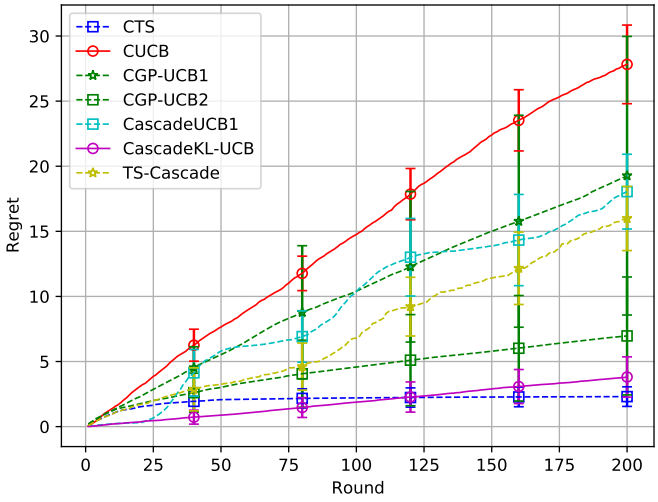

Figure 3: Regrets for the disjunctive cascading bandit problem, when the expected base-arm outcomes are sampled randomly. (L=40, R=5, K=5).

Table 1: Mean value and standard deviations of ComGP-UCB's regret after $T = 200$ rounds with different length-scale parameters. The expected base-arm outcomes are sampled from a GP with squared exponential kernel with variance set to 1 and length-scale set to 0.8. Note that this a separate experiment from the ones whose results plotted before, only to compare the effect of length-scale parameter on the performance of ComGP-UCB. The kernel is same for all four setups, which is the squared-exponential kernel (L= 200, R=1, K=10).

Algorithm	**Length-scale**	**Regret** ($T = 200$)
ComGP-UCB	0.8	131.73 ± 0.47
ComGP-UCB	0.6	164.52 ± 0.37
ComGP-UCB	0.4	204.44 ± 0.43
ComGP-UCB	0.2	283.98 ± 0.62

A.2.2 A Synthetic Problem

We consider a synthetic problem where the reward of a super-arm S is simply the sum of its base arms' individual rewards. Formally, $R(S(t), \boldsymbol{Y}(t)) = \sum_{i \in S(t)} \bar{r}_i^t$, where $\bar{r}_i^t$ denotes the outcome of base arm $i \in [m]$ at round t.

In Figure 4, we have the regrets of the algorithms when the expected base-arm outcomes are sampled from a GP with squared-exponential kernel [1]. We see that in this case, ComGP-UCB significantly outperforms the other state-of-the-art algorithms. In Figure 5, we investigate the setting where multiple users a served simultaneously. Again, we see that ComGP-UCB outperforms the other algorithms. Similar to before, in Figure 6, we observe that the ComGP-UCB does not significantly improve other algorithms results when the expected base-arm outcomes are sampled randomly. Also, we again note that the ComGP-UCB with a lower length-scale parameter (ComGP-UCB2) works better. Please see Table 1 to observe the effect of the length-scale parameter on the performance of ComGP-UCB. In that setup, the expected base-arm outcomes are sampled from a GP with a squared exponential kernel with $ls = 0.8$. We can observe that the best performance is obtained when the learner's (GP) length-scale parameter is matching the actual length-scale parameter used for sampling the expected base-arm outcomes. As expected, the performance degrades as the learner's length-scale parameter decreases (as the learner starts assuming weaker dependence between similar base-arms).

[1] Variance set to 1 and length-scale set to 0.8

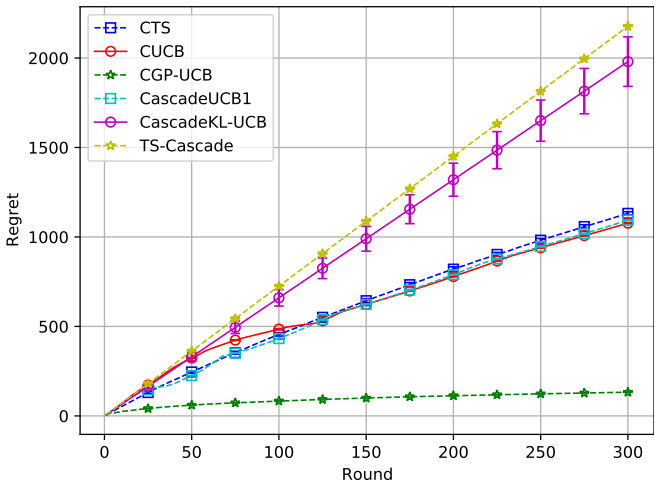

Figure 4: Regrets for the synthetics linear bandit problem, when there is high correlation between expected base-arm outcomes (L=300, R=1, K=10. Variance parameter is set to 1, and length-scale parameter is set to 0.8 for Squared-Exponential Kernel. Likelihood variance is set to 0.0001).

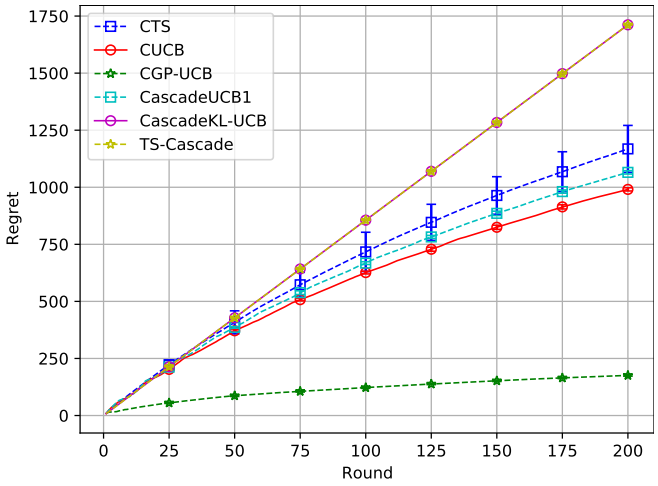

Figure 5: Regrets for the synthetic linear bandit problem, when there is high correlation between expected base-arm outcomes (L=40, R=5, K=5. Variance parameter is set to 1, and length-scale parameter is set to 0.8 for Squared-Exponential Kernel. Likelihood variance is set to 0.0001).

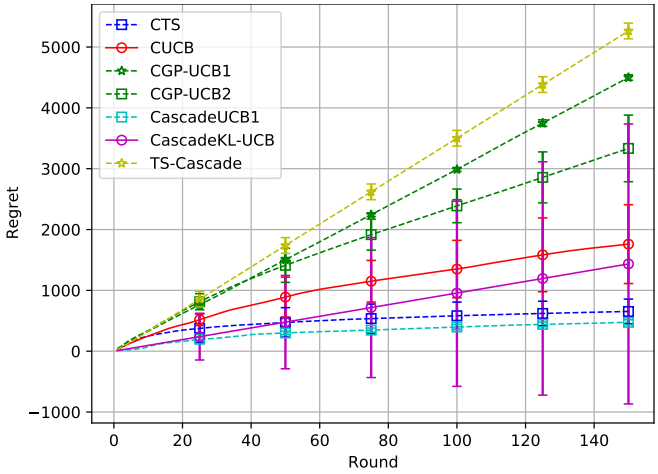

Figure 6: Regrets for the synthetic linear bandit problem, when the expected base-arm outcomes are sampled randomly (L=30, R=5, K=5).

References

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.

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.