Learning-Aided Control for Constrained Contextual Bandits

Paper ID: 000

Abstract

Confidence Set based Lyapunov Optimization.

1 Introduction (Outline)

Contextual bandits

The contextual bandit problem (Langford and Zhang 2007; Lu, Pál, and Pál 2010; Zhou 2015) is an important extension of the classic multi-armed bandit (MAB) problem (Auer, Cesa-Bianchi, and Fischer 2002), where the agent can observe a set of features, referred to as *context*, before making a decision. After the random arrival of a context, the agent chooses an action and receives a random reward with expectation depending on both the context and action. To maximize the total reward, the agent needs to make a careful tradeoff between taking the best action based on the historical performance (exploitation) and discovering the potentially better alternative actions under a given context (exploration). This model has attracted much attention as it fits the personalized service requirement in many applications such as clinical trials, online recommendation, and online hiring in crowdsourcing. Existing works try to reduce the regret of contextual bandits by leveraging the structure of the context-reward models such as linearity (Li et al. 2010) or similarity (Slivkins 2014), and more recent work (Agarwal et al. 2014) focuses on computationally efficient algorithms with minimum regret. For Markovian context arrivals, algorithms such as UCRL (Auer and Ortner 2007) for more general reinforcement learning problem can be used to achieve logarithmic regret.

Bandit with Budget Constraints

However, traditional contextual bandit models do not capture an important characteristic of real systems: in addition to time, there is usually a cost associated with the resource consumed by each action and the total cost is limited by a budget in many applications. Taking crowdsourcing (Badanidiyuru, Kleinberg, and Singer 2012) as an example, the budget constraint for a given set of tasks will limit the number of workers that an employer can hire. Another example is the clinical trials (Lai and Liao 2012), where each treatment is usually costly and the budget of a trial is limited. Although budget constraints have been studied in non-contextual bandits where logarithmic or sublinear regret is achieved (Tran-Thanh et al. 2012; Badanidiyuru, Kleinberg, and Slivkins 2013; Jiang and Srikant 2013; Slivkins 2013; Xia et al. 2015; Combes, Jiang, and Srikant 2015; Flajolet and Jaillet 2015), as we will see later, these results are inapplicable in the case with observable contexts.

Constrained contextual bandits

The RCB paper (Badanidiyuru, Langford, and Slivkins 2014): show $O(\sqrt{T})$ regret is achievable; but restrict to a finite policy set, and the proposed algorithm is computationally inefficient;

Our UCB-ALP paper (Wu et al. 2015): proposed a computationally efficient algorithm, achieve logarithmic regret; focus on single budget constraint, fixed cost.

The two technical reports:

(Agrawal, Devanur, and Li 2015): propose a computationally efficient algorithm; again, restrict to a finite policy set;

(Agrawal and Devanur 2015): By combining the online learning algorithm and confidence bound techniques, (Agrawal and Devanur 2015) proposes a computationally efficient algorithm for linear contextual bandits with global concave objective and convex budget constraints. They first solve the feasibility problem and then convert the original constrained reward maximization problem by introducing an estimate of the optimal reward. This requires a lot of effort on estimating the optimal reward. Further, for the linear reward and constrained case, it can only deal with the case where $B = \Omega(T^{3/4})$, it is still an open problem when $B = \Theta(\sqrt{T})$.

Lyapunov Optimization (Backpressure/MaxWeight):

Lyapunov optimization is a method of using a Lyapunov function to optimally control a dynamical system. This method has been extensively studied in queueing networks (**?**). Earlier work focuses on the stability of queueing systems by minimizing the Lyapunov drift. Combination of Lyapunov drift and the sum of penalty leads to the drift-plus-penalty algorithm for joint network stability and long-term average penalty minimization (). The drift-plus-penalty procedure can also be used to compute solutions to linear programming and convex optimization (). Traditional Lyapunov optimization assume the knowledge of expected rewards and costs.

When the reward and cost as well as context distribution are unknown statistics, some papers [Neely], [Ouyang], [Huang2015MobiHoc] propose ϵ-first, or epoch approach to achieve near optimal performance under average constraints. Unlike bandit literature, these papers does not provide analysis on the cumulative regret.

This work:

In this paper, we study contextual bandits with general budget constraints. We consider very general settings with unknown context distribution, multiple budget constraints, and random cost. We propose a learning-aided control algorithm, referred to as *Confidence Lyapunov Optimization*, which combines UCB with Lyapunov optimization method. This algorithm is computationally efficient where in each round one just needs to sort the indices of K actions (following (Wu et al. 2015) will need to solve an LP problem with JK variables; The RCB paper (Badanidiyuru, Langford, and Slivkins 2014) is difficult to implement) and select the action with highest value.

We would like to obtain theoretical bounds for the cumulative regret. We show that for $B = \Theta(T)$ this may be weaker than (Badanidiyuru, Langford, and Slivkins 2014) and (Agrawal and Devanur 2015), where they just need $B = \Omega(\sqrt{T})$, the proposed algorithm achieves $O(\sqrt{T \log T})$ regret /*Wu: A critical issue: the UCB/LCB may be correlated with the virtual queue; In literature: (Ouyang et al. 2010) and (Agrawal and Devanur 2015) have similar issue but they don't appropriately address it; (Neely, Rager, and La Porta 2012) tried to address this issue by independently sampling the historic observations and consider the delayed estimates.*/.

Furthermore, using the framework of Lyapunov optimization, the proposed algorithm can also be extended to Markovian case (so is (Agrawal and Devanur 2015)).

Compared to (Agrawal and Devanur 2015), our proposed CLO algorithm directly use the actual evolution of budget constraint, and can converge much faster.

2 Constrained Contextual Bandits

We consider a contextual bandit problem with multiple budget constraints. At each time-slot t, a context $X_t \in \mathcal{X}$ arrives, independently following the identical distribution. For simplicity, we consider discrete and finite contexts, where $\mathcal{X} = \{1, 2, \ldots, J\}$ and $\mathbb{P}\{X_t = j\} = \pi_j$. Upon observing the context, the agent takes an action $A_t \in \mathcal{A} = \{0\} \cup \{1, 2, \ldots, K\}$, where "0" represents a *dummy* action that the agent skips the context. When taking an action $k \in \{1, 2, \ldots, K\}$, the agent will receive a reward Y_t and consume $Z_{i,t}$ amount of type-i resource, where $1 \leq i \leq M$ and M is the number of types of resource. We consider finite reward and cost, and thus we can normalize the reward and cost such as $Y_t \in [0, 1]$ and $Z_{i,t} \in [0, 1]$ for all t and i. Conditioned on the context and the taken action, the expected reward is given by $\mathbb{E}[Y_t | X_t = j, A_t = k] = u_{j,k}$ and the expected cost of type-i resource is $\mathbb{E}[Z_{i,t} | X_t = j, A_t = k] = c_{j,k}^{(i)}$. Note that the system statistics, π_j's, $u_{j,k}$'s, and $c_{j,k}^{(i)}$'s are unknown to the agent.

A constrained contextual bandit algorithm Γ decides which action to take given the historic observations and the current states. The objective of Γ is to maximize the expected total reward subject to the budget constraints, i.e.,

$$\max_{\Gamma} \; U_\Gamma(T, \boldsymbol{b}) = \sum_{t=1}^{T} \mathbb{E}[Y_t], \tag{1}$$

$$\text{s. t.} \; \sum_{t=1}^{T} \mathbb{E}[Z_{i,t}] \leq b_i T, 1 \leq i \leq M, \tag{2}$$

where the expectation operation is taken over the realization of contexts, actions, rewards and costs; and $b_i T$ is the budget associated with each type i of resource. We first consider soft budget constraints, and discuss the hard constraint later (Section **??**).

Let $U^*(T, \boldsymbol{b})$ be expected total reward achieved by the oracle, i.e., the optimal algorithm with known system statistics. The regret of an algorithm Γ is defined as the gap between its performance and the oracle, i.e.,

$$R_\Gamma(T, \boldsymbol{b}) = U^*(T, \boldsymbol{b}) - U_\Gamma(T, \boldsymbol{b}). \tag{3}$$

3 Confidence Lyapunov Optimization

In this section, we propose a Confidence Lyapunov Optimization (CLO) for constrained contextual bandits. We first discuss the oracle solutions with known system statistics, and the approximate solution when the context distribution is unknown. Then, we propose our CLO algorithm when the system statistics are unknown.

3.1 Oracle Solution: Static Linear Programming

When the system statistics, including π_j's, $u_{j,k}$'s, and $c_{j,k}^{(i)}$'s, are known, the oracle solution can take a randomized policy and make the optimal decision by solving a static linear programming (SLP) problem, where the optimal decision only depends on the current context X_t. Specifically, let $p_{j,k}$ be the probability that an algorithm takes action k under context j. Then, the optimal solution can be obtained by solving the following linear programming problem:

$$\max_{p_{j,k}} \; \sum_{j=1}^{J} \sum_{k=1}^{K} p_{j,k} \pi_j u_{j,k} \tag{4}$$

$$\text{s. t.} \; \sum_{j=1}^{J} \sum_{k=1}^{K} p_{j,k} \pi_j c_{j,k}^{(i)} \leq b_i, \;\; \forall i, \tag{5}$$

$$\sum_{k=1}^{K} p_{j,k} \leq 1, \;\; \forall j. \tag{6}$$

Let $v^*(\boldsymbol{b})$ be the optimal solution of the above problem. Then $U^*(T, \boldsymbol{b}) = T v^*(\boldsymbol{b})$ is the performance of the oracle solution when the time horizon is T. As shown in (Wu et al. 2015), if we consider the hard constraints as (Badanidiyuru, Langford, and Slivkins 2014; Wu et al. 2015), then $U^*(T, \boldsymbol{b})$ provides an upper bound for the oracle solution.

3.2 Approximate Solution with Unknown Context Distribution

When $u_{j,k}$'s and $c_{j,k}^{(i)}$'s are known, but π_j's are unknown, a Lyapunov optimization algorithm (Neely 2010) can be applied and has been shown to be ϵ-optimal. The key idea of this Lyapunov optimization algorithm is to estimate the Lagrangian multipliers of the SLP problem and make decisions based on these estimates.

Specifically, when the statistics are known, consider the Lagrangian of the SLP problem,

$$\begin{aligned}&\mathcal{L}(\boldsymbol{p},\boldsymbol{\gamma})\\&= V\sum_{j=1}^{J}\sum_{k=1}^{K} p_{j,k}\pi_j u_{j,k} + \sum_{i=1}^{M}\gamma_i\sum_{j=1}^{J}\big[\sum_{k=1}^{K} p_{j,k}\pi_j c_{j,k}^{(i)} - b_i\big]\\&= \sum_{j=1}^{J}\Big\{V\sum_{k=1}^{K} p_{j,k}\pi_j u_{j,k} + \sum_{i=1}^{M}\gamma_i\big[\sum_{k=1}^{K} p_{j,k}\pi_j c_{j,k}^{(i)} - b_i\big]\Big\}.\end{aligned} \tag{7}$$

Note that adding a positive parameter V here will not change the optimal solution $\boldsymbol{p}$, but it will play an important role in balancing between the reward and constraints later when the system statistics is unknown.

When the context distribution $\boldsymbol{\pi}$ is unknown, the Lyapunov optimization algorithm maintains virtual queues corresponding to each type of resource constraints, which will serve as the Lagrangian multipliers when making decision. Specifically, consider a virtual queue $Q_i(t)$ for each type of resource i, which evolves as follows:

$$Q_i(t+1) = [Q_i(t) - b_i]^+ + Z_{i,t}, \tag{8}$$

where $[x]^+ = \max\{x, 0\}$.

At slot t, the Lyapunov optimization algorithm chooses the decision based on $X_t = j$ and $Q_i(t)$'s:

$$(\mathcal{P}_{\mathrm{LO}})\ \max_{k\in\mathcal{A}}\ Vu_{j,k} + \sum_{i=1}^{M} Q_i(t)[b_i - c_{j,k}^{(i)}]$$

The Lyapunov optimization algorithm is shown to achieve ϵ-optimality (Neely 2010), i.e., $U_{\mathrm{Lyp}}(T,\boldsymbol{b}) \geq U^*(T,\boldsymbol{b}) - O(\frac{1}{V})$, and $\mathbb{E}[Q_i(T)] = O(V)$ represents the budget violation.

3.3 Confidence Lyapunov Optimization

When the system statistics are unknown, we propose a Confidence Lyapunov Optimization (CLO) algorithm, where we implement Lyapunov optimization algorithm within the confidence bounds of the reward $u_{j,k}$ and the cost $c_{j,k}^{(i)}$.

Specifically, let $N_{j,k}(t-1)$ be the number of rounds that action k has been taken under context j until time-slot $t-1$, i.e., $N_{j,k}(t-1) = \sum_{\tau=1}^{t-1}\mathbb{1}(X_\tau = j, A_\tau = k)$. For $t > 0$ and $N_{j,k}(t-1) > 0$, the empirical value of the reward and cost are given as

$$\bar{u}_{j,k}(t) = \frac{1}{N_{j,k}(t-1)}\sum_{\tau=1}^{t-1} Y_\tau \mathbb{1}(X_\tau = j, A_\tau = k),$$

$$\bar{c}_{j,k}^{(i)}(t) = \frac{1}{N_{j,k}(t-1)}\sum_{\tau=1}^{t-1} Z_{i,\tau} \mathbb{1}(X_\tau = j, A_\tau = k).$$

Then, we define the confidence bounds of the expected reward as $\check{u}_{j,k}(t) = \bar{u}_{j,k}(t) - \sqrt{\frac{\alpha\log t}{N_{j,k}(t-1)}}$ and $\hat{u}_{j,k}(t) = \bar{u}_{j,k}(t) + \sqrt{\frac{\alpha\log t}{N_{j,k}(t-1)}}$, and the confidence bounds of the expected cost as $\check{c}_{j,k}^{(i)}(t) = \bar{c}_{j,k}^{(i)}(t) - \sqrt{\frac{\alpha\log t}{N_{j,k}(t-1)}}$ and $\hat{c}_{j,k}^{(i)}(t) = \bar{c}_{j,k}^{(i)}(t) + \sqrt{\frac{\alpha\log t}{N_{j,k}(t-1)}}$, where $\alpha > 0.5$ is a constant. For $N_{j,k}(t-1) = 0$, we let $\check{u}_{j,k}(t) = 0$, $\hat{u}_{j,k}(t) = 1$, $\check{c}_{j,k}^{(i)}(t) = 0$, and $\hat{c}_{j,k}^{(i)}(t) = 1$.

Using the Chernoff-Hoeffding bounds, we have the following lemma:

Lemma 1. *For any $t > 0$,*

$$\mathbb{P}\{u_{j,k} \in [\check{u}_{j,k}(t), \hat{u}_{j,k}(t)]\} \geq 1 - \frac{2}{t^{2\alpha}},$$

$$\mathbb{P}\{c_{j,k} \in [\check{c}_{j,k}(t), \hat{c}_{j,k}(t)]\} \geq 1 - \frac{2}{t^{2\alpha}}.$$

Now we present the CLO algorithm, as shown in Algorithm 1. The idea of CLO is to solve the Lyapunov optimization problem $(\mathcal{P}_{\mathrm{LO}})$ within the confidence bounds, i.e., with additional conditions $u_{j,k}(t) \in [\check{u}_{j,k}(t), \hat{u}_{j,k}(t)]$ and $c_{j,k}^{(i)}(t) \in [\check{c}_{j,k}^{(i)}(t), \hat{c}_{j,k}^{(i)}(t)]$. For constrained contextual bandits with linear reward and budget constraints, we can easily that we just need to replace $u_{j,k}$ with its UCB $\hat{u}_{j,k}(t)$ and $c_{j,k}$ with its LCB $\check{c}_{j,k}(t)$, respectively, as described in Eq. (9).

Algorithm 1 Confidence Lyapunov Optimization (CLO)

Input: T, $\boldsymbol{b}$, V;
Init: $N_{j,k}(0) \leftarrow 0$; $\bar{u}_{j,k}(0) = 0$, $\bar{c}_{j,k}^{(i)}(0) = 0$;
for $t = 1$ **to** T **do**
  Observe context X_t;
  Calculate UCB $\hat{u}_{X_t,k}(t)$ and LCB $\check{c}_{X_t,k}(t)$;
  Choose the action as follows

$$A_t = \arg\max_{k\in\mathcal{A}} V\hat{u}_{j,k}(t) - \sum_{i=1}^{M} Q_i(t)\check{c}_{j,k}^{(i)}(t) \tag{9}$$

  Receive reward Y_t and costs $Z_{i,t}$'s,
  and update the empirical reward and costs.
end for

3.4 Regret Analysis

In this section, we study the regret of CLO. Although the intuition behind the CLO algorithm is natural, the analysis of its regret is challenging. The coupling effect in CLO significantly increases the analysis complexity compared to traditional Lyapunov optimization algorithm and the UCB algorithm for unconstrained bandits. Specifically, the values of $\hat{u}_{j,k}(t)$'s and $\check{c}_{j,k}(t)$'s will affect the the decision and then the evolution of $Q_i(t)$. Conversely, the values of $Q_i(t)$'s affect the decision and determine which action to explore. Moreover, in principle, the value of $Q_i(t)$ is unbounded, and thus certain tiny error of $\hat{u}_{j,k}(t)$ or $\check{c}_{j,k}(t)$ may be amplified

significantly and cause a wrong decision. This is unlike the dual variables in (Agrawal and Devanur 2015), which are updated within a bounded set. We address these difficulties by studying the cumulative impact of estimation errors and the large deviation properties of virtual queues. We show that the CLO algorithm achieves $O(\sqrt{T\log T})$ regret.

Theorem 1. *Given $V=\sqrt{T}$, the regret of CLO satisfies*

$$R_{\text{CLO}}(T)=O(\sqrt{T\log T}). \tag{10}$$

To bound the regret of CLO, we need to bound the number of error decisions. Under CLO, there are two sources for these errors, one is the estimation errors for $u_{j,k}$ and $c^{(i)}_{j,k}$, the other is the deviation of the virtual queues. We bound these two types of errors, respectively.

When the rewards $u_{j,k}$'s and the costs $c_{j,k}$'s are unknown, the error decision is inevitable when the action k has not been taken under context j sufficiently. Given a constant $\delta>0$, we define $F^{(\delta)}_{j,k}(T)$ as the number of time-slots that the *executed while under-sampling* event occurs, i.e.,

$$F^{(\delta)}_{j,k}(T)=\sum_{t=1}^{T}\mathbb{1}(X_t=j,A_t=k,N_{j,k}(t-1)\le\frac{\alpha\log T}{\delta^2}). \tag{11}$$

According to the definition of $F^{(\delta)}_{j,k}(T)$, we can easily see that $F^{(\delta)}_{j,k}(T)\le\frac{\alpha\log T}{\delta^2}$. Thus, consider all j and k, we have

Lemma 2. *The sum of $F^{(\delta)}_{j,k}(T)$ is bounded as follows:*

$$\sum_{j,k}F^{(\delta)}_{j,k}(T)=\frac{\alpha JK\log T}{\delta^2}. \tag{12}$$

Based on Lemma 2 and according to the properties of confidence bounds, we can bound the deviation of the virtual queues as follows:

Lemma 3. *Under CLO,*

$$\mathbb{P}\{Q_i(t)\ge(2\delta)^{-1}+3/2)V+\kappa\log V\}\le\eta e^{-\kappa\log V},\ \forall t, \tag{13}$$

where η is a constant.

Discussions: one issue in the proof - dependency of $\hat{u}$, $\check{c}$ and $Q(t)$. The drift is calculated conditioned on $\boldsymbol{Q}(t)$. However, the UCB/LCB may depend on $\boldsymbol{Q}(t)$.

Proof. Define a new Lyapunov function $L_0(t)=\frac{1}{2}||\boldsymbol{Q}(t)||^2$, and its drift as $\Delta_0(t)=L_0(t+1)-L_0(t)$. Then, the drift of $L_0(t)$ under CLO satisfies

$$\begin{aligned}&\mathbb{E}[\Delta_0(t)|\boldsymbol{Q}(t)]\\ =\ &\frac{1}{2}\sum_{i=1}^{M}\left\{\mathbb{E}\left[[Q_i(t)-b_i]^+ + Z_i(t)\middle|\boldsymbol{Q}(t)\right]^2-Q_i^2(t)\right\}\\ \le\ &\sum_{i=1}^{M}Q_i(t)\mathbb{E}[Z_i(t)-b_i|\boldsymbol{Q}(t)].\end{aligned} \tag{14}$$

Under CLO, we know that

$$\mathbb{E}_{X_t\sim\mathcal{D}_X}[V\hat{u}_{X_t,A_t}(t)-\sum_{i=1}^{M}\check{c}^{(i)}_{X_t,A_t}(t)Q_i(t)]\ge 0. \tag{15}$$

On the other hand, let $\delta_{j,k}(t)=\sqrt{\frac{\alpha\log t}{N_{j,k}(t-1)}}$. Then, with probability $(1-\frac{2JK}{t^{2\alpha}})$, $u_{X_t,A_t}+2\delta_{X_t,A_t}(t)\ge\hat{u}_{X_t,A_t}(t)$, and $c^{(i)}_{X_t,A_t}-2\delta_{X_t,A_t}(t)\le\check{c}^{(i)}_{X_t,A_t}(t)$. Thus,

$$\begin{aligned}&\mathbb{E}_{X_t\sim\mathcal{D}_X}[Vu_{X_t,A_t}-\sum_{i=1}^{M}c^{(i)}_{X_t,A_t}Q_i(t)\\ &\qquad+2\delta_{X_t,A_t}(t)(V+\sum_{i=1}^{M}Q_i(t))]\\ \ge\ &\mathbb{E}_{X_t\sim\mathcal{D}_X}[V\hat{u}_{X_t,A_t}(t)-\sum_{i=1}^{M}\check{c}^{(i)}_{X_t,A_t}(t)Q_i(t)]\ge 0.\end{aligned} \tag{16}$$

According to Lemma 2, we know that in at least $T-\frac{\alpha JK\log T}{\delta^2}$ slots, we have $N_{X_t,A_t}(t-1)>\frac{\alpha\log T}{\delta^2}$ and thus $\delta_{X_t,A_t}(t)\le\delta$. Consequently,

$$\begin{aligned}\mathbb{E}[\Delta_0(t)|\boldsymbol{Q}(t)]\ &\le\ \mathbb{E}_{X_t\sim\mathcal{D}_X}[Vu_{X_t,A_t}]\\ &\quad+2\delta(V+\sum_{i=1}^{M}Q_i(t))-\sum_{i=1}^{M}Q_i(t)b_i\\ &\le\ (1+2\delta)V+\sum_{i=1}^{M}Q_i(t)(2\delta-b_i)\end{aligned} \tag{17}$$

Let $b_{\min}=\min_i b_i$, and $\delta=b_{\min}/4$. Then when $L_0(t)$ exceeds certain threshold, it will have a negative drift, i.e.,

$$\begin{aligned}&\mathbb{E}[\Delta_0(t)|L_0(t)\ge((2\delta)^{-1}+3/2)V)^2]\\ \le\ &(1+2\delta)V-(1+3\delta)V\le-\delta V.\end{aligned} \tag{18}$$

Therefore, we have $L_0(t)\le((2\delta)^{-1}+3/2)V)^2$ with high probability according to [Hajek]. Thus, $Q_i(t)\le\sqrt{2}(2\delta)^{-1}+3/2)V$ with high probability. □

Now we bound the cumulative regret of the single-step objective function, which determine the evolution of the Lyapunov function and thus the total reward and costs. We note that the traditional Lyapunov optimization algorithm maximize the following function in each time slot:

$$\Phi_t(j,k)=Vu_{j,k}-\sum_{i=1}^{M}Q_i(t)c^{(i)}_{j,k}, \tag{19}$$

while CLO maximize the following function based on the confidence bounds:

$$\hat{\Phi}_t(j,k)=V\hat{u}_{j,k}-\sum_{i=1}^{M}Q_i(t)\check{c}^{(i)}_{j,k}. \tag{20}$$

The following lemma state that under CLO, the cumulative regret for this single-step objective function is bounded by $O(V\sqrt{T\log T})$.

Lemma 4. *Under CLO,*

$$\sum_{t=1}^{T}\mathbb{E}[\hat{\Phi}_t(X_t,A_t)-\Phi_t(X_t,A_t)]\le O(V\sqrt{T\log T}) \tag{21}$$

Proof. /*Wu: Some constants depend on Lemma 3 and will be determined later.*/

For each t, if $X_t = j$, $A_t = k$, we have

$$\begin{aligned}&\hat{\Phi}_t(j,k) - \Phi_t(j,k)\\ =\;& V[\hat{u}_{j,k}(t) - u_{j,k}] - \sum_{i=1}^{M} Q_i(t)[\check{c}_{j,k}^{(i)}(t) - c_{j,k}^{(i)}] \quad (22)\end{aligned}$$

Note that $Q_i(t) \leq ?V$, we have

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}[\hat{\Phi}_t(X_t,A_t) - \Phi_t(X_t,A_t)]\\ \leq\;& V\sum_{t=1}^{T}\mathbb{E}[\hat{u}_{X_t,A_t}(t) - u_{X_t,A_t}]\\ &-?V\sum_{i=1}^{M}\sum_{t=1}^{T}[\check{c}_{X_t,A_t}^{(i)}(t) - c_{X_t,A_t}^{(i)}]. \quad (23)\end{aligned}$$

Similar to [Agrawal2014EC], we have

$$\mathbb{E}\big[\sum_{t=1}^{T}[\hat{u}_{X_t,A_t}(t) - u_{X_t,A_t}]\big] \leq O(\sqrt{JKT\log T}), \quad (24)$$

and

$$-\mathbb{E}\big[\sum_{t=1}^{T}[\check{c}_{X_t,A_t}^{(i)}(t) - c_{X_t,A_t}^{(i)}]\big] \leq O(\sqrt{JKT\log T}). \quad (25)$$

The conclusion then follows.

Proof. (**Proof of Theorem 1**).

Step 1: Single-slot Lyapunov-drift

$$\begin{aligned}\Delta_V(t) &= L(t+1) - L(t) - VY_t \quad (26)\\ &\leq G - \{Vu_{j,A_t} + \sum_{i=1}^{M} Q_i(t)[b_i - c_{j,A_t}^{(i)}]\}\\ &= G - \sum_{i=1}^{M} Q_i(t)b_i - \{Vu_{j,A_t} - \sum_{i=1}^{M} Q_i(t)c_{j,A_t}^{(i)}\}\\ &\quad (27)\end{aligned}$$

where $G = ??$.

We have

$$\begin{aligned}\Delta_V(t) &\leq G - \sum_{i=1}^{M} Q_i(t)B_i/T - \{Vu_{j,A_t} - \sum_{i=1}^{M} Q_i(t)c_{j,A_t}^{(i)}\}\\ &= G - \sum_{i=1}^{M} Q_i(t)B_i/T - \hat{\Phi}_t(j,A_t)\\ &\quad +[\hat{\Phi}_t(j,A_t) - \Phi_t(j,A_t)] \quad (28)\end{aligned}$$

Under CLO, for $\hat{\Phi}_t(j,A_t)$, we have (p^* is the optimal solution for the fixed LP problem)

$$\begin{aligned}&\mathbb{E}_{X_t\sim\mathcal{D}_x}[\hat{\Phi}_t(X_t,A_t)|\boldsymbol{Q}(t)]\\ \geq\;& \mathbb{E}_{X_t\sim\mathcal{D}_x,A_t\sim p^*}[\hat{\Phi}_t(X_t,A_t)|\boldsymbol{Q}(t)]\\ \geq\;& \mathbb{E}_{X_t\sim\mathcal{D}_x,A_t\sim p^*}[\Phi_t(j,A_t)|\boldsymbol{Q}(t)] \text{ (with prob. } 1-JK/t)\\ \geq\;& \hat{v}(T,B) - \sum_{i=1}^{M}(B_i/T - \epsilon_i)Q_i(t). \quad (29)\end{aligned}$$

Thus, with probability $1 - JK/t$, we have

$$\begin{aligned}\mathbb{E}[\Delta_V(t)] &= \mathbb{E}[L(t+1) - L(t) - VY_t]\\ &\leq G - \hat{v}(T,B)V - \sum_{i=1}^{M}\epsilon_i\mathbb{E}[Q_i(t)]\\ &\quad +\mathbb{E}[\hat{\Phi}_t(X_t,A_t) - \Phi_t(X_t,A_t)]. \quad (30)\end{aligned}$$

Taking the telescoping sum over $t = 1, 2, \ldots, T$ we have

$$\begin{aligned}&\mathbb{E}[L(T+1)] - \mathbb{E}[L(1)] - V\mathbb{E}[\hat{U}_{\mathrm{CLO}}(T,B)]\\ \leq\;& GT - TV\hat{v}(T,B) - \sum_{t=1}^{T}\sum_{i=1}^{M}\epsilon_i\mathbb{E}[Q_i(t)]\\ &+\sum_{t=1}^{T}\mathbb{E}[\hat{\Phi}_t(X_t,A_t) - \Phi_t(X_t,A_t)]. \quad (31)\end{aligned}$$

Here we let $\hat{U}_{\mathrm{CLO}}(T,B)$ be the total reward under CLO from slot 1 to T. Note that this is not the real reward of CLO because the algorithm will terminate when one of the resource has been exhausted. We will bound the gap $hatU_{\mathrm{CLO}}(T,B) - U_{\mathrm{CLO}}(T,B)$ latter. Thus,

$$\begin{aligned}&\mathbb{E}[\hat{U}_{\mathrm{CLO}}(T,B)]\\ &\geq T\hat{v}(T,B) - \frac{GT}{V} + \frac{\mathbb{E}[L(T+1)] - \mathbb{E}[L(1)]}{V}\\ &+\frac{\sum_{t=1}^{T}\sum_{i=1}^{M}\epsilon_i\mathbb{E}[Q_i(t)]}{V} - \frac{\sum_{t=1}^{T}\mathbb{E}[\hat{\Phi}_t(X_t,A_t) - \Phi_t(X_t,A_t)]}{V}\\ &\geq \hat{U}(T,B) - G\sqrt{T} - \frac{\sum_{t=1}^{T}\mathbb{E}[\hat{\Phi}_t(X_t,A_t) - \Phi_t(X_t,A_t)]}{V}.\\ &\quad (32)\end{aligned}$$

From Lemma 4, we have $\sum_{t=1}^{T}\mathbb{E}[\hat{\Phi}_t(X_t,A_t) - \Phi_t(X_t,A_t)] = VO(\sqrt{T\log T})$. The conclusion then follows.

4 Simulation Results

5 Discussions

5.1 Thompson-Lyapunov Optimization

Combination of Thompson Sampling and Lyapunov Optimization.

Consider Bernoulli reward and cost with expectation $u_{j,k}$ and $\bar{c}_{j,k}^{(i)}$.

For general case, one can use the method in (Agrawal and Goyal 2012) to convert the problem to a binary scenario.

6 Conclusion

References

Agarwal, A.; Hsu, D.; Kale, S.; Langford, J.; Li, L.; and Schapire, R. E. 2014. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning (ICML)*.

Agrawal, S., and Devanur, N. R. 2015. Linear contextual bandits with global constraints and objective. *arXiv preprint arXiv:1507.06738*.

Algorithm 2 Thompson-Lyapunov Optimization (TLO)

Input: T, $\boldsymbol{b}$, V;

Init: $S^u_{j,k}(0) \leftarrow 0$, $F^u_{j,k}(0) \leftarrow 0$; $S^{c(i)}_{j,k}(0) \leftarrow 0$, $F^{c(i)}_{j,k}(0) \leftarrow 0$;

for $t = 1$ **to** T **do**

Observe context X_t;

Draw samples from the posterior:

$$\theta^u_{j,k}(t) \sim \text{Beta}(S^u_{j,k}(t-1)+1, F^u_{j,k}(t-1)+1),$$
$$\theta^{c(i)}_{j,k}(t) \sim \text{Beta}(S^{c(i)}_{j,k}(t-1)+1, F^{c(i)}_{j,k}(t-1)+1).$$

Choose the action as follows

$$A_t = \arg\max_{k\in\mathcal{A}} V\theta^u_{j,k}(t) - \sum_{i=1}^{M} Q_i(t)\theta^{c(i)}_{j,k}(t) \quad (33)$$

Receive reward Y_t and costs $Z_{i,t}$, and update the empirical reward and costs.

end for

Agrawal, S., and Goyal, N. 2012. Analysis of Thompson Sampling for the multi-armed bandit problem. In *Conference on Learning Theory (COLT)*.

Agrawal, S.; Devanur, N. R.; and Li, L. 2015. Contextual bandits with global constraints and objective. *arXiv preprint arXiv:1506.03374*.

Auer, P., and Ortner, R. 2007. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems (NIPS)*, 49–56.

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

Badanidiyuru, A.; Kleinberg, R.; and Singer, Y. 2012. Learning on a budget: posted price mechanisms for online procurement. In *ACM Conference on Electronic Commerce*, 128–145.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2013. Bandits with knapsacks. In *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, 207–216.

Badanidiyuru, A.; Langford, J.; and Slivkins, A. 2014. Resourceful contextual bandits. In *Conference on Learning Theory (COLT)*.

Combes, R.; Jiang, C.; and Srikant, R. 2015. Bandits with budgets: Regret lower bounds and optimal algorithms. In *ACM Sigmetrics*.

Flajolet, A., and Jaillet, P. 2015. Low regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800*.

Jiang, C., and Srikant, R. 2013. Bandits with budgets. In *IEEE 52nd Annual Conference on Decision and Control (CDC)*, 5345–5350.

Lai, T. L., and Liao, O. Y.-W. 2012. Efficient adaptive randomization and stopping rules in multi-arm clinical trials for testing a new treatment. *Sequential Analysis* 31(4):441–457.

Langford, J., and Zhang, T. 2007. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 817–824.

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A contextual-bandit approach to personalized news article recommendation. In *ACM International Conference on World Wide Web (WWW)*, 661–670.

Lu, T.; Pál, D.; and Pál, M. 2010. Contextual multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, 485–492.

Neely, M. J.; Rager, S. T.; and La Porta, T. F. 2012. Max weight learning algorithms for scheduling in unknown environments. *IEEE Trans. on Automatic Control* 57(5):1179–1191.

Neely, M. J. 2010. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks* 3(1):1–211.

Ouyang, W.; Murugesan, S.; Eryilmaz, A.; and Shroff, N. B. 2010. Scheduling with rate adaptation under incomplete knowledge of channel/estimator statistics. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, 670–677. IEEE.

Slivkins, A. 2013. Dynamic ad allocation: Bandits with budgets. *arXiv preprint arXiv:1306.0155*.

Slivkins, A. 2014. Contextual bandits with similarity information. *The Journal of Machine Learning Research* 15(1):2533–2568.

Tran-Thanh, L.; Chapman, A. C.; Rogers, A.; and Jennings, N. R. 2012. Knapsack based optimal policies for budget-limited multi-armed bandits. In *AAAI Conference on Artificial Intelligence*.

Wu, H.; Srikant, R.; Liu, X.; and Jiang, C. 2015. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. In *The 29th Annual Conference on Neural Information Processing Systems (NIPS)*.

Xia, Y.; Li, H.; Qin, T.; Yu, N.; and Liu, T.-Y. 2015. Thompson sampling for budgeted multi-armed bandits. In *International Joint Conference on Artificial Intelligence*.

Zhou, L. 2015. A survey on contextual multi-armed bandits. *arXiv preprint arXiv:1508.03326*.