IEOR 8100-001: Learning and Optimization for Sequential Decision Making 03/07/16

Lecture 14: Bandits with Budget Constraints

Instructor: Shipra Agrawal Scribed by: Zhipeng Liu

1 Problem definition

In the regular Multi-armed Bandit problem, we pull an arm I_t at each round t and a reward r_t is collected, which depend on the bandit I_t. Now suppose that each round after pulling the arm, a cost c_t is also incurred. When the total cost till time t surpass a given level B, the algorithm stops. This setting of problem is called Bandits with constraints. The formal description is following.

At time t,

- pull an arm $I_t = i$, observe reward $r_t \in [0,1]$ and cost $\mathbf{c_t} \in [\mathbf{0},\mathbf{1}]^{\mathbf{d}}$.
- Given $I_t = i$, $(r_t, \mathbf{c}_t) \sim D_i$, where D_i is a joint distribution dependent on arm i, denote $E[r_t|I_t = i] = \mu_i$, $E[c_{tj}|I_t = i] = C_{ij}$.
- Stop when any budget constraint is violated.
- Goal is to

$$\begin{aligned} \text{maximize} \quad & \sum_{t=1}^{T} r_t \\ \text{subject to} \quad & \sum_{t} \mathbf{c_t} \leq \mathbf{B}, \end{aligned}$$

here, $\mathbf{B} \geq \mathbf{0}$.

Remark. The above setting is also called **Bandit with Knapsacks** (BwK). Usually we replace the constraint with a simplified version by letting $B = \min_j B_j$, and

$$\sum_t c_{tj} \leq B, \quad j = 1, \ldots, d.$$

Example 1. *Dynamic pricing with limited supply*
Suppose at price q_i, the product has probability $S(q_i)$ to be sold, observing revenue q_i and incurring inventory decreasing by 1. So the distribution of reward and cost for the bandit problem is: pulling arm i,

$$(r_t, c_t) = \begin{cases} (q_i, 1) & w.p.\ S(q_i), \\ (0,0) & w.p.\ 1 - S(q_i) \end{cases}$$

2 Optimal static policy

A policy is a mapping which maps history to action. It can be pure static, in which case we keep pulling the single arm with highest expected reward. It can be mixed, in which case we pull arms randomly according to some distribution. Or it can be dynamic, where each turn we make decision based on the history and remaining resource. In this section we will show that there is an optimal static policy which has expected reward at least optimal dynamic and it satisfies constraints in expectation.

Suppose each turn we pull arm i with probability p_i, and the constraints are all satisfied in expectation. Then the optimal total reward (OPT) is;

$$\begin{aligned} \max \quad & \sum_i p_i \mu_i T \\ \text{subject to} \quad & \sum_i p_i C_{ij} \le \frac{B}{T}, \qquad j = 1, \dots, d \\ & \sum_i p_i \le 1. \end{aligned}$$

Notice that the second constraint implies that we allow not pulling any arm during one round.

Next we prove (OPT) is better than what a optimal dynamic policy could achieve.
Define X_i: total number of times an optimal policy picked arm i, $\tilde{p}_i = \frac{E[X_i]}{T}$.
Any feasible policy must satisfy

$$\sum_{s=1}^{t} c_{sj} \le B \quad \forall t = 1, \dots, T.$$

Take expectations on both sides and we have

$$\begin{aligned} & E[\sum_{i=1,\dots,N} \sum_{t:I_t=i} c_t] \le B \\ \iff \quad & \sum_i E[X_i C_{ij}] \le B \\ \iff \quad & \sum_i (\tilde{p}_i T) C_{ij} \le B. \end{aligned}$$

Therefore $\{\tilde{p}_i\}_{i=1}^N$ is feasible to (OPT), and the total expected reward of this optimal policy

$$\begin{aligned} E[\sum_i X_i \mu_i] &= \sum_i E[X_i \mu_i] \\ &= \sum_i T \tilde{p}_i \mu_i \\ &\le OPT. \end{aligned}$$

3 Bandit algorithm: reducing to unconstrained f problem

Suppose the algorithm stops at time τ when a budget constraint is violated. Define the regret of such policy

$$R(T) = OPT - E[\sum_{t=1}^{T} r_t].$$

Since the total reward of optimal dynamic policy is bounded by OPT, the gap between optimal dynamic policy and the algorithm is bounded by $R(T)$.
Now we reduce the constrained problem to an unconstrained problem with nonlinear objective function by applying Lagrangian multiplier to the constrained problem. The new unconstrained problem will maximize $f\left(\sum_{t=1}^{T} \frac{r_t}{T}, \sum_{t=1}^{T} \frac{\mathbf{c}_t}{T}\right)$,

which is a concave and Lipschitz continuous function.

$$f\left(\sum_{t=1}^{T}\frac{r_t}{T},\sum_{t=1}^{T}\frac{\mathbf{c}_t}{T}\right)=\sum_{t=1}^{T}\frac{r_t}{T}-z\max_{j=1,\ldots,d}\left(\frac{c_{tj}}{T}-\frac{B}{T},0\right).$$

In the above definition of $f(\cdot)$, The first item is average reward, and the second item is the penalty from the maximum violation of budget constraint. We define the penalty coefficient z as

$$z=\frac{2OPT}{B},$$

which we will explain momentarily.

Suppose we relax the budget to $B+\Delta B$, and define $\epsilon=\frac{\Delta B}{B}$. Denote the (OPT) with budget $B(1+\epsilon)$ as $\text{OPT}^{1+\epsilon}$. Any $\mathbf{p}$ $\text{OPT}^{1+\epsilon}$-feasible, since

$$T\frac{\sum_i p_i}{1+\epsilon}C_{ij}\le\frac{B(1+\epsilon)}{1+\epsilon}\quad\forall j=1,\ldots,d$$

therefore $\frac{\mathbf{p}}{1+\epsilon}$ is OPT-feasible. Thus we have

$$\begin{aligned}\sum_{i=1}^{N}p_i\mu_i&=\sum_{i=1}^{N}\frac{p_i}{1+\epsilon}\mu_i(1+\epsilon)\\&\le(1+\epsilon)\text{OPT},\end{aligned}$$

so if we relax the budget constraint by ΔB, the optimal value of OPT will increase by no more than $\frac{\Delta B}{B}\text{OPT}$. So we can set $z=\frac{2\text{OPT}}{B}$ which guarantees that violating the budget won't give benefit in terms of increasing value of $f(\cdot)$. Theorem 2 below will provide the exact relation between optimizing f and the Bandits with Knapsacks problem. The significance of this value of z will be more precisely illustrated in the proof of that theorem.

Now we define the regret of the algorithm with objective function $f(\cdot)$,

$$R_f(T)=\text{OPT}_f-TE\left[f\left(\frac{\sum_{t=1}^{T}r_t}{T},\frac{\sum_{t=1}^{T}\mathbf{c_t}}{T}\right)\right],$$

where

$$\text{OPT}_f=\max_{p:\sum_i p_i\le 1}Tf(\sum_i p_i\mu_i,\sum_i p_iC_i)=\sum_i p_i\mu_i-Z\max_j(C_{ij}-\frac{B}{T},0)$$

And we define $R'(T)$ as the regret of the constrained problem with larger budget

$$B'=B+\frac{2}{z}R_f(T)+\tilde{O}(\sqrt{B})$$

$$R'(T)=\text{OPT}'-TE\left[\sum_{t=1}^{\tau}r_t\right].$$

where τ is the first time a budget B' is violated. We conclude this lecture with the following theorem.

Theorem 2. *If an algorithm achieves $R_f(T)$ regret for unconstrained f problem, then*

$$R'(T)\le 3R_f(T)+z\tilde{O}(\sqrt{B}),$$

and the algorithm will not violate B' at any time step $t\le T$ with high probability.

Proof. (**Proof outline:**) The proof follows from the following two claims:

1. $\text{OPT}_f \geq \text{OPT}' - z(B' - B)$.
2. Let $\mathbf{c}_t$ be the cost of the decision at time t for unconstrained f algorithm, then, with high probability, $\sum_{t=1}^{T} c_{tj} \leq B'$ for all j.

If above two claims are true, then $\tau = T$ with high probability, and

$$R'(T) = \text{OPT}' - \mathbb{E}[\sum_{t=1}^{\tau} r_t] \leq \text{OPT}_f - \mathbb{E}[\sum_{t=1}^{T} r_t] = R_f(T) + z \max_{j=1,\ldots,d} \left(\frac{c_{tj}}{T} - \frac{B}{T}, 0 \right) \leq R_f(T) + z(B' - B),$$

Using $B' - B = \frac{2R_f(T)}{z} + \tilde{O}(\sqrt{B})$, we get the theorem statement. Next, we prove the above two claims.

The first claim holds because the optimal solution (in fact any solution p such that $\sum_i p_i \leq 1, p_i \geq 0$) for OPT' forms a feasible solution for OPT_f, with value $\text{OPT}' - z(B' - B)$.

For second claim, let M be the maximum budget violation above B' by the algorithm for the unconstrained f problem, i.e.,

$$M := \max_{j=1,\ldots,d} \left(\frac{c_{tj}}{T} - \frac{B}{T}, 0 \right)$$

Then,

$$\begin{aligned} f\left(\frac{\sum_{t=1}^{T} r_t}{T}, \frac{\sum_{t=1}^{T} \mathbf{c_t}}{T} \right) &= E\left[\frac{\sum_{t=1}^{T} r_t}{T} \right] - zM \\ &\leq \frac{OPT}{T} + \frac{zM}{2} - zM \\ &\leq \frac{OPT_f}{T} - \frac{zM}{2} \end{aligned}$$

The second last inequality follows using our earlier observation that $\text{OPT}^{1+\epsilon} \leq \text{OPT}(1+\epsilon)$, and $z = \frac{2\text{OPT}}{B}$. Last inequality follows because $\text{OPT}_f \geq \text{OPT}$ (the optimal solution p for OPT forms a feasible solution for OPT_f, with value OPT). Then, rearranging,

$$TM \leq \frac{2}{z} \left(\text{OPT}_f - f\left(\frac{\sum_{t=1}^{T} r_t}{T}, \frac{\sum_{t=1}^{T} \mathbf{c_t}}{T} \right) \right) = \frac{2}{z} R_f(T)$$

Therefore, by definition of M,

$$\mathbb{E}[\sum_t c_{tj}] \leq B + \frac{2R_f(T)}{z}$$

So, that (using Azuma-Hoeffding), with high probability $\sum_{t=1}^{T} c_{tj} \leq B + \frac{2}{Z} R_f(T) + \tilde{O}(\sqrt{B}) = B'$, proving the second claim. □

Why is above theorem useful?

Conceptually, above theorem shows that to bound regret of Bandits with knapsacks, we can instead use an algorithm that has small regret bound for the unconstrained f bandit problem. In next lecture we will prove regret bounds for the unconstrained f problem. In particular, for the f discussed above, a uniform bound on regret of $R_f(T) \leq \tilde{O}(z\sqrt{NB})$ can be proven (though we will not prove this special case in class). Then, given above theorem,

we could solve the unconstrained f bandit problem with slightly smaller budget $\bar{B} = B - \frac{2}{z}\tilde{O}(\sqrt{NB}) - \tilde{O}\sqrt{B}$, to get regret

$$R_f(T) \leq \tilde{O}(\bar{z}\sqrt{N\bar{B}}) \leq \tilde{O}(z\sqrt{NB}\frac{B}{\bar{B}}) \leq \tilde{O}(z\sqrt{NB})$$

so that above theorem will give

$$R(T) \leq 3R_f(T) + z\tilde{O}\sqrt{B} \leq \tilde{O}(z\sqrt{NB} + z\tilde{O}(\sqrt{B}))$$

Using $z = \frac{2\text{OPT}}{B}$, we get a bound of $\tilde{O}(\frac{\text{OPT}}{\sqrt{B}})$ on regret for Bandits with knapsacks, i.e. multiplicative guarantee that the reward achieved by bandit algorithm is at least $\text{OPT}(1 - \tilde{O}(\frac{1}{\sqrt{B}}))$.

Note that above algorithm assumed that the value of OPT is known (OPT was used to set the value of z). If OPT is not known, then it needs to be estimated using pure exploration in the beginning. This initial exploration might cause additional regret. This is a limitation of the above approach of reducing the problem to unconstrained f problem. For a direct approach without knowing OPT, refer to Section 4.2 of [1].

References

[1] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006. ACM, 2014.