IEOR 8100-001: Learning and Optimization for Sequential Decision Making 02/24/16

Lecture 12: Contextual bandits

Instructor: Shipra Agrawal Scribed by: Chaoxu Zhou

1 Review

Let X be the space of context, $A = \{1, \ldots, N\}$ be the space of all possible actions. When $x_t \in X$ is observed, we take an action $a_t \in A$ based on the observation and get the reward $r_t(a_t) \in [0, 1]$. In stochastic settings, we assume that (x_t, r_t) independently and identically follows an unknown distribution $\mathcal{D}$. Moreover, assume there is a space $\Pi = \{\pi \,|\, \pi : X \to A\}$ of policies and we want to construct algorithms that maximize

$$\mathbb{E}\{R(T)\} = \sum_{t=1}^{T} \mathbb{E}[r(a_t)] - \max_{\pi \in \Pi} \mathbb{E}[r(\pi(x))]. \tag{1}$$

In last lecture, we talked about two types of contextual bandit algorithms, namely

Exp 4 algorithm : General, but computationally not feasible.

ϵ-greedy algorithm : Only relevant to stochastic problems with *i.i.d.* assumptions. Computationally feasible.

ϵ-greedy algorithm chooses the best policy so far and randomized uniformly over the rests and Exp 4 maintains a distribution of good coverage over policies. More specifically, when there are several policies have almost the same rewards as the best one, ϵ-greedy algorithm will choose the best one with a very high probability while Exp 4 tends to choose not only the best one but also those 'good' ones in a more uniformly manner.

2 Improve the exploration strategy

The high level idea of the algorithm we want to construct is the following:

Step 1 Construct distribution P'_t overt Π,

Step 2 Sample $\pi_t \sim P'_t$,

Step 3 Play $\pi_t(x_t)$ and observe $r_t(\pi_t(x_t))$,

Step 4 Update P'_t to P'_{t+1}.

Exp 4 algorithm has the same structure but with a very specific choice of the distribution P_t. We want to construct a distribution that is computationally efficient and has good exploration property as Exp 4. Ideally, we want to construct exploration distribution only support on those 'good' policies so that have low regrets. However, in practice we don't know which are the ones have low regrets. Then we will construct distribution that only plays low empirical regret policies. Let

$$\hat{r}_t(a) = \frac{r_t(a)\mathbb{I}\{a = a_t\}}{P_t(a)}, \tag{2}$$

where $P_t(a) = \sum_{\pi(x_t)=a, \pi \in \Pi} P'_t(\pi)$. The empirical regret can be otained by replacing $\mathbb{E}r_t(a_t)$ by $\hat{r}_t(a)$ in (1). (**Remark:** In MAB problems, we applied strategies such as UCB, Exp 3 and ϵ-greedy algorithms to explore arms

other than the one with the lowest empirical regret in order to obtain an optimal regret bound. Similarly, in contextual bandit problems, only using the policy with the lowest empirical regret can not be optimal.) To bound the difference between the empirical mean and population mean, the following concentration inequalities may be needed.

Lemma 1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables and each $X_i \in [a_i, b_i]$, then*

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}X_i| \geq t) \leq 2\exp\big\{-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\big\}.$$

Lemma 2 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent real-valued random variables that $\mathbb{E}X_i = 0$ and $X_i \leq 1$. Let $\sigma^2 = \sum_{i=1}^{n} var(X_i)/n$. Then for any $t > 0$,*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} X_i > t) \leq \exp\big\{-\frac{nt^2}{2(\sigma^2 + t/3)}\big\}$$

Note that in order to obtain a high probability bound of the difference between empirical mean and population mean, we need the information of the ranges (in Hoeffding's inequality) and variances (in Bernstein's inequality) of $\hat{r}_t(\pi(x_t))$. From (2), the expression of $\hat{r}_t$, we know that when $\mathbb{P}(a)$ is close to zero, $\hat{r}_t$ is large. Therefore in order to upper bound $\hat{r}_t$ (0 is a natural lower bound of $\hat{r}_t$), we need to impose the condition that

$$\max_{a\in A} P_t(a) \geq \mu_t > 0.$$

Next, we need to have an upper bound of the variance of $\hat{r}_t$. For every $\pi \in \Pi$,

$$\begin{aligned}
\mathrm{var}\{\hat{r}_t(\pi(x_t))\} &\leq \mathbb{E}\{\hat{r}_t(\pi(x_t))\}^2 \\
&= \mathbb{E}\frac{r_t^2(\pi(x_t)\mathbb{I}\{\pi(x_t) = a\}}{P_t^2(\pi(x_t))} && \text{(defintion of } \hat{r}_t\text{)} \\
&\leq \mathbb{E}\frac{\mathbb{I}\{\pi(x_t) = a\}}{P_t^2(\pi(x_t))} && (0 \leq r_t \leq 1) \\
&= \mathbb{E}\frac{1}{P_t(\pi(x_t))} \\
&\equiv b_t(\pi).
\end{aligned}$$

If we can make both the range and variance of $\hat{r}_t$ small, then the empirical regret will be close to expected regret with high probability for every $\pi \in \Pi$. We can use the following feasibility problem to quantify the above argument:

$$\begin{aligned}
&\text{Find } \ P' \text{ over } \Pi && \text{(OP)} \\
&\text{such that } \ \sum_{\pi\in\Pi} P'(\pi)\hat{\mathrm{Regret}}_t(\pi) \leq C\sqrt{\frac{N\log|\Pi|}{\sigma t}} \\
&\qquad \forall \pi \in \Pi, \mathbb{E}\frac{1}{P(\pi(x))} \leq b_t(\pi) \equiv 2N + \frac{1}{\mu_t}\hat{\mathrm{Regret}}(\pi)
\end{aligned}$$

The first constraint can be interpreted as exploitation of the the past data to have a small empirical regret. The second constraint can be interpreted as exploration of polices other than the best one. The first term in the bound of the second constraint can be interpreted as uniform of the 'good' sub-optimal policies and second term in the bound can be interpreted as the price we need to pay for the bad polices. Note that the second constraint involves an expectation with respect to an unknown distribution $\mathcal{D}$. We can use sample average to replace it and the result won't be changed. With solutions to the feasibility problem, we have the following almost optimal statistical property

Theorem 3. *If P_t satisfies (OP) for all $t = 1, \ldots, T$ and $\pi_t \sim P_t$, then*

$$\sum_{t=1}^{T} \mathbb{E} r(\pi_t(x_t)) \geq \max_{\pi \in \Pi} T\mathbb{E}[r(\pi(a))] - O(\sqrt{\frac{\tilde{NT} \log |\Pi|}{\delta}})$$

with probability at least $1 - \delta$.

It can be proved that (OP) is a convex programming problem. However, there are $|\Pi|$ many constraints which can be exponentially large. The basic idea of solving (OP) is to move some probability mass on to the policy that violate the constraints most. It can be shown that (OP) can be solved iteratively by calling a single maximization oracle. In $O(\sqrt{T/(N \log |\Pi|)})$ calls of the maximization oracle, all constraints in (OP) are satisfied. Therefore, in contrast to Exp 4 algorithm, the algorithm based on solving (OP) is computationally tractable. In applications, we can use a warm start (use the distribution that satisfies all constraints in the last step as a starting point to solve (OP in current step) to have a better performance.