Lecture 20

Instructor: Haipeng Luo

1 Toward Optimal and Efficient Contextual Bandit

We have discussed Epsilon-Greedy last time, an oracle-efficient but suboptimal algorithm for the i.i.d. contextual bandit problem. In this lecture we discuss an optimal but inefficient algorithm. It conveys important ideas based on which one can further derive both optimal and efficient algorithm that we will discuss next time.

Let's first recall what the key difficulty is in getting the optimal regret. A very reasonable template for an algorithm is to come up with a distribution over policies P_t at time t and then pick a_t according to $P_t(\cdot|x_t)$ but with a small amount of uniform exploration. To this end, define $P_t^\mu(\cdot|x_t)$ to be the mixture of $P_t(\cdot|x_t)$ and some uniform exploration so that

$$P_t^\mu(a|x_t) = (1 - K\mu)P_t(a|x_t) + \mu$$

for some parameter $\mu \le 1/K$. Also recall the notation $\bar{\ell}(\pi) = \mathbb{E}_{(x,\ell)\sim\mathcal{D}}[\ell(\pi(x))]$ for the expected loss of a policy π and $\bar{\ell}_t(\pi) = \frac{1}{t}\sum_{\tau=1}^t \widehat{\ell}_\tau(\pi(x_\tau))$ for the empirical average loss where $\widehat{\ell}_\tau$ is the usual importance weighted estimators. The most important part of analyzing such algorithms is to understand the concentration of $\bar{\ell}_t(\pi)$. As shown before, the conditional variance of $\widehat{\ell}_t(\pi) - \bar{\ell}_t(\pi)$ is bounded as

$$\mathbb{E}_{x_t,\ell_t,a_t}\left[\left(\widehat{\ell}_t(\pi) - \bar{\ell}(\pi)\right)^2\right] \le \mathbb{E}_{x_t,\ell_t,a_t}\left[\widehat{\ell}_t(\pi)^2\right] = \mathbb{E}_{x_t,\ell_t}\left[\frac{\ell_t(\pi)^2}{P_t^\mu(\pi(x_t)|x_t)}\right] \le \mathbb{E}_{x_t}\left[\frac{1}{P_t^\mu(\pi(x_t)|x_t)}\right].$$

Define for a distribution P and a policy π (and implicitly a marginal distribution over the contexts and a parameter μ)

$$V(P,\pi) = \mathbb{E}_x\left[\frac{1}{P^\mu(\pi(x)|x)}\right],$$

so that the conditional variance is simply bounded by $V(P_t,\pi)$. By Freedman's inequality, we have with probability at least $1-\delta$,

$$|\bar{\ell}_t(\pi) - \bar{\ell}(\pi)| \le \mathcal{O}\left(\sqrt{\left(\frac{1}{t}\sum_{\tau=1}^t V(P_\tau,\pi)\right)\frac{\ln(1/\delta)}{t}} + \frac{\ln(1/\delta)}{\mu t}\right).$$

For Epsilon-Greedy, we simply bound each $V(P_\tau,\pi)$ by $1/\mu$, which then leads to $O(T^{\frac{2}{3}})$ regret. If we could ensure that $V(P_\tau,\pi)$ is much smaller, say a constant that is independent of T, then there is hope in getting $\mathcal{O}(\sqrt{T})$ regret (this is indeed the case in the full information case).

To get a sense of how small this quantity can be, one can first consider the special case where there is only one possible context x. In this case the policies are naturally grouped into (at most) K classes according to which action they pick given x. Then simply picking a distribution P over these policies such that $P(\cdot|x)$ is uniform would make $V(P,\pi) \le K$ for any π.

When there are many different contexts, the argument above does not generalize. However, the conclusion turns out to be still true up a factor of two, as shown by the following lemma.

Lemma 1. *For any policy space Π, any context distribution, and any $\mu \le 1/K$, there always exists a distribution $P \in \Delta(\Pi)$ such that $V(P,\pi) \le 2K$ for all $\pi \in \Pi$.*

Proof. It is clear that the statement is equivalent to the following:

$$\min_{P\in\Delta(\Pi)} \max_{\pi\in\Pi} V(P,\pi) \le 2K.$$

We thus work on the minimax expression on the left. If $\mu \ge 1/(2K)$, then the statement trivially holds since $V(P,\pi) \le 1/\mu$. Below we assume $\mu \le 1/(2K)$ and thus $1 - K\mu \ge 1/2$. First note that we can "linearize" the maximization part because maximization over the simplex can always be achieved by one of the elements:

$$\min_{P\in\Delta(\Pi)} \max_{\pi\in\Pi} V(P,\pi) = \min_{P\in\Delta(\Pi)} \max_{Q\in\Delta(\Pi)} \mathbb{E}_{\pi\sim Q}[V(P,\pi)].$$

Next we apply Sion's minimax theorem to swap the min and max

$$\min_{P\in\Delta(\Pi)} \max_{Q\in\Delta(\Pi)} \mathbb{E}_{\pi\sim Q}[V(P,\pi)] = \max_{Q\in\Delta(\Pi)} \min_{P\in\Delta(\Pi)} \mathbb{E}_{\pi\sim Q}[V(P,\pi)].$$

By picking a specific $P = Q$, the last expression is clearly bounded by $\max_{Q\in\Delta(\Pi)} \mathbb{E}_{\pi\sim Q}[V(Q,\pi)]$. Now note that

$$\begin{aligned}\mathbb{E}_{\pi\sim Q}[V(Q,\pi)] &= \mathbb{E}_x\left[\sum_{\pi\in\Pi} \frac{Q(\pi)}{Q^\mu(\pi(x)|x)}\right] \le \mathbb{E}_x\left[\sum_{\pi\in\Pi} \frac{Q(\pi)}{(1-K\mu)Q(\pi(x)|x)}\right] \\ &\le 2\mathbb{E}_x\left[\sum_{\pi\in\Pi} \frac{Q(\pi)}{Q(\pi(x)|x)}\right] = 2\mathbb{E}_x\left[\sum_{a=1}^{K} \sum_{\pi:\pi(x)=a} \frac{Q(\pi)}{Q(a|x)}\right] = 2K,\end{aligned}$$

which completes the proof. □

This lemma shows that we can always find a distribution that leads to low variance for the loss estimators. In other words, the distribution does a good job in terms of exploration. However, such a distribution says nothing about exploitation. Indeed, it is even independent of the observed losses.

One way to address this issue is to ensure that we only keep around "good" policies. Specifically, we start with the whole policy space $\Pi_1 = \Pi$; at each time t, we find a distribution P_t over Π_t that induces low variance (Lemma 1 shows that it always exists no matter what Π_t is); finally we remove all bad policies in Π_t based on what we have observed and obtain a new policy space Π_{t+1}. This final step can be done by simply checking how much worse a policy is in terms of the empirical performance compared to the empirically best policy, since we know that empirical data concentrates well to the truth due to the low variance of estimators. This algorithm is called Policy Elimination [Dudík et al., 2011] and is shown in Algorithm 1.

As mentioned this is not an efficient algorithm. Moreover, since the definition of V depends on the the unknown context distribution, it does not even seem to be a valid algorithm. However, the latter issue can be solved by simply replacing the context distribution by the empirical distribution, that is, a uniform distribution over $x_1, \ldots, x_t$ at time t. The analysis remains similar except that one more step of concentration is needed now. For simplicity, we will skip this step and assume that the context distribution is known. The following theorem shows that Policy Elimination achieves the optimal regret (recall that regret is defined as $\mathcal{R}_T = \sum_{t=1}^{T}(\ell_t(a_t) - \bar{\ell}(\pi^\star))$.

Theorem 1. *Policy Elimination ensures* $\mathcal{R}_T = \widetilde{\mathcal{O}}\left(\sqrt{TK\ln(N/\delta)} + K\ln(N/\delta)\right)$ *with probability at least* $1-\delta$.

Proof. Clearly we have $\Pi_T \subset \cdots \subset \Pi_1 = \Pi$ and thus for any $\pi \in \Pi_t$ we have $V(P_\tau,\pi) \le 2K$ for any $\tau = 1, \ldots, t$ by the algorithm. Therefore, by Freedman's inequality and union bound, we have with probability $1-\delta/2$, for all $t \in [T]$ and all $\pi \in \Pi_t$,

$$\begin{aligned}|\bar{\ell}_t(\pi) - \bar{\ell}(\pi)| &\le 2\sqrt{\left(\frac{1}{t}\sum_{\tau=1}^{t} V(P_\tau,\pi)\right)\frac{\ln(4NT/\delta)}{t}} + \frac{\ln(4NT/\delta)}{\mu t} \\ &\le 2\sqrt{\frac{2K\ln(4NT/\delta)}{t}} + \frac{\ln(4NT/\delta)}{\mu t} = \frac{\epsilon_t}{2}. \qquad (1)\end{aligned}$$

Algorithm 1: Policy Elimination

```
Input: failure probability $\delta \in (0,1)$
Initialization: let $\Pi_1 = \Pi$, $\epsilon_t = 4\sqrt{\frac{2K\ln(4NT/\delta)}{t}} + \frac{2\ln(4NT/\delta)}{\mu t}$, $\mu = \min\left\{\frac{1}{K}, \sqrt{\frac{\ln(TN/\delta)\ln T}{TK}}\right\}$
for $t = 1, \ldots, T$ do
    find $P_t$ such that $V(P_t, \pi) \le 2K$ for all $\pi \in \Pi_t$
    play $a_t \sim P_t^\mu(\cdot|x_t)$
    update $\Pi_{t+1} = \left\{\pi \in \Pi_t : \bar{\ell}_t(\pi) \le \bar{\ell}_t(\pi_t^\star) + \epsilon_t\right\}$ where $\pi_t^\star = \operatorname{argmin}_{\pi\in\Pi_t} \bar{\ell}_t(\pi)$
```

Conditioning on this event, we can show that $\pi^\star$ is never removed from the policy space: inductively assuming $\pi^\star \in \Pi_t$, we have

$$\bar{\ell}_t(\pi^\star) \le \bar{\ell}(\pi^\star) + \frac{\epsilon_t}{2} \le \bar{\ell}(\pi_t^\star) + \frac{\epsilon_t}{2} \le \bar{\ell}_t(\pi_t^\star) + \epsilon_t,$$

which means $\pi^\star$ will stay in Π_{t+1}. Finally, applying Azuma inequality we have with probability $1-\delta$,

$$\begin{aligned}
\sum_{t=1}^T \ell_t(a_t) &\le \sum_{t=1}^T \mathbb{E}_{x_t,\ell_t,a_t}[\ell_t(a_t)] + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) \\
&\le \sum_{t=1}^T \mathbb{E}_{x_t,\ell_t}\left[\sum_{\pi\in\Pi_t} P_t(\pi)\ell_t(\pi(x_t))\right] + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) \\
&= \sum_{t=1}^T \mathbb{E}_{\pi\sim P_t}[\bar{\ell}(\pi)] + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) \\
&\le \sum_{t=1}^T \mathbb{E}_{\pi\sim P_t}[\bar{\ell}_{t-1}(\pi)] + \frac{1}{2}\sum_{t=2}^T \epsilon_{t-1} + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) && \text{(by Eq. (1))} \\
&\le \sum_{t=1}^T \bar{\ell}_{t-1}(\pi^\star) + \frac{3}{2}\sum_{t=2}^T \epsilon_{t-1} + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) && \text{(since } \pi \in \Pi_t\text{)} \\
&\le \sum_{t=1}^T \bar{\ell}(\pi^\star) + 2\sum_{t=2}^T \epsilon_{t-1} + TK\mu + \mathcal{O}\left(\sqrt{T\ln(1/\delta)}\right) && \text{(by Eq. (1))} \\
&\le \sum_{t=1}^T \bar{\ell}(\pi^\star) + \mathcal{O}\left(\sqrt{TK\ln(TN/\delta)} + \frac{\ln(TN/\delta)\ln T}{\mu} + TK\mu\right),
\end{aligned}$$

which competes the proof with the optimal tuning of μ. □

References

Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2011.