STATS 710 – Seq. Dec. Making with mHealth Applications Oct 13, 2016

Lecture 11: Epoch Greedy Algorithm

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Peng Liao*

1 Recap from Lecture 10

Recall from last lecture that the general flow of information for contextual bandits is

1: **for** $t = 1$ to T **do**
2: Learner sees $X_t \in \mathcal{X}$
3: Learner selects action $A_t \in \mathcal{A}$
4: Learner receives reward $R_t = R_t^{A_t}$
5: **end for**

In the stochastic setting, we assume that the context-reward pair $(X_t, R_t) = \{X_t, R_t^a, a \in \mathcal{A}\}$ are i.i.d. across time with distribution D and the expected regret against a policy class Π is defined as

$$\mathcal{R}_T(\mathcal{L}, D, \Pi) = \sup_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=1}^{T} R_t^{\pi(X_t)} - \sum_{t=1}^{T} R_t^{A_t}\right] = TV(\pi^*) - \mathbb{E}\left[\sum_{t=1}^{T} R_t^{A_t}\right]$$

where $V(\pi) := \mathbb{E}_{(X,R)\sim D}\left[R_t^{\pi(X)}\right]$ is the value function and $\pi^* = \text{argmax}_{\pi\in\Pi} V(\pi)$ is the optimal policy. We showed that when the context space $\mathcal{X}$ is finite, by using the idea of reduction to multiple MAB the regret bound scales as $\sqrt{K \log K}\sqrt{T|\mathcal{X}|}$ for EXP3. The dependence on $\sqrt{|\mathcal{X}|}$ cannot be avoided. Note that this result is only valid when the context space is finite and we are not making any assumptions on how different contexts relate to the expected reward. In general, we have two options to bypass this finiteness assumption:

Option 1. Do not model $\mathbb{E}[R^a|X]$, but try to compete with a class of policies.

Option 2. Model $\mathbb{E}[R^a|X]$ and compete with the best policy overall.

2 Epoch Greedy Algorithm [LZ08]

We will start with a simple algorithm in which we assume that the total number of rounds is known upfront and prove the regret bound of order $(\text{VC-dim}(\Pi))^{1/3}T^{2/3}$. Then we will introduce the Epoch Greedy algorithm that generalizes this simple algorithm: it does not require the knowledge of T, but could still achieve the same rate if increasing the size of epoch appropriately. We will use the VC dimension in the proof to upper bound the expected sup norm of empirical process over the policy class. This is a standard result in statistics [VDVW96] and machine learning (see the lecture notes in the email). In the following, for simplicity we always assume action are binary, i.e. $\mathcal{A} = \{0, 1\}$, and the policies are deterministic.

2.1 A Simple Algorithm with known T.

Algorithm 1 Simplified Version of Epoch Greedy Algorithm

Input: The total rounds T. Rounds of exploration T_0. Policy Class Π.
for $t = 1$ to T_0 **do**
  Learner sees $X_t \in \mathcal{X}$
  Learner selects action A_t uniformly over $\mathcal{A}$ (e.g. $A_t \sim$ Bernoulli(0.5))
  Learner receives reward $R_t = R_t^{A_t}$
end for
Compute $\hat{\pi} = \arg\max_{\pi \in \Pi} \frac{1}{T_0} \sum_{t=1}^{T_0} 2R_t^{A_t} \mathbf{1}_{\{\pi(X_t)=A_t\}}$
for $t = T_0 + 1$ to T **do**
  Learner sees $X_t \in \mathcal{X}$
  Learner selects action $A_t = \hat{\pi}(X_t)$
  Learner receives reward $R_t = R_t^{A_t}$
end for

Theorem 1. *Suppose $R_t^a \in [0,1]$ and VC-dim$(\Pi) < \infty$. Then the expected regret of Algorithm 1 at round T with $T_0 < T$ exploration rounds is bounded by*

$$\mathcal{R}_T \leq T_0 + CT\sqrt{\frac{\textit{VC-dim}(\Pi)}{T_0}}$$

where C is some universal constant. If T_0 is chosen optimally, i.e. $T_0 = (CT)^{2/3}(\textit{VC-dim}(\Pi))^{1/3}$, then $\mathcal{R}_T = O((\textit{VC-dim}(\Pi))^{1/3}T^{2/3})$.

Proof of Theorem 1. Since rewards are bounded within $[0,1]$, the regret incurred in first T_0 rounds are at most T_0. For the $T - T_0$ exploitation rounds, the expected regret can be expressed as

$$\begin{aligned}
(T-T_0)V(\pi^*) - \mathbb{E}\Big[\sum_{t=T_0+1}^{T} R_t^{A_t}\Big] &= (T-T_0)V(\pi^*) - \mathbb{E}\Big[\sum_{t=T_0+1}^{T} R_t^{\hat{\pi}(X_t)}\Big] \\
&= (T-T_0)V(\pi^*) - \mathbb{E}\Big[\mathbb{E}\Big[\sum_{t=T_0+1}^{T} R_t^{\hat{\pi}(X_t)}\Big]\Big|H_{T_0}\Big] \\
&= (T-T_0)\mathbb{E}\left[V(\pi^*) - V(\hat{\pi})\right]
\end{aligned}$$

where $H_{T_0} := \{X_i, A_i, R_i\}_{i=1}^{T_0}$ is the history up to time T_0. Denote by $\hat{V}(\pi) = \frac{1}{T_0}\sum_{t=1}^{T_0} 2R_t^{A_t}\mathbf{1}_{\{\pi(X_t)=A_t\}}$ the estimated value of policy π after T_0 exploration rounds. It is straightforward to see that $\hat{V}(\pi)$ is unbiased: $\mathbb{E}[\hat{V}(\pi)] = V(\pi)$. Using the definition of $\hat{\pi}$, i.e. $\hat{V}(\hat{\pi}) = \text{argmax}_{\pi\in\Pi}\hat{V}(\pi)$, we have

$$\begin{aligned}
\mathbb{E}\left[V(\pi^*) - V(\hat{\pi})\right] &= \mathbb{E}\left[V(\pi^*) - \hat{V}(\pi^*) + \hat{V}(\pi^*) - \hat{V}(\hat{\pi}) + \hat{V}(\hat{\pi}) - V(\hat{\pi})\right] &(1)\\
&\leq \mathbb{E}\left[V(\pi^*) - \hat{V}(\pi^*) + \hat{V}(\hat{\pi}) - V(\hat{\pi})\right] &(2)\\
&\leq 2\mathbb{E}\left[\sup_{\pi\in\Pi}\left|V(\pi) - \hat{V}(\pi)\right|\right] \leq C\sqrt{\frac{\text{VC-dim}(\Pi)}{T_0}} &(3)
\end{aligned}$$

Thus the total regret $\mathcal{R}_T \le T_0 + C(T-T_0)\sqrt{\frac{\text{VC-dim}(\Pi)}{T_0}} \le T_0 + CT\sqrt{\frac{\text{VC-dim}(\Pi)}{T_0}}$.

□

2.2 Epoch Greedy Algorithm

Algorithm 2 Epoch Greedy Algorithm

Input: Policy Class Π. l_j = the number of exploitation rounds in the j-th epoch, $j = 1, 2 \ldots$.
Initialize $D_0 = \emptyset$ and $t_1 = 1$
for epoch $j = 1, 2, \ldots$ **do**
  $t = t_j$
  Learner sees $X_t \in \mathcal{X}$
  Learner selects action A_t uniformly in $\mathcal{A}$
  Learner receives reward $R_t = R_t^{A_t}$
  $D_j = D_{j-1} \cup \{X_t, A_t, R_t\}$
  Compute $\hat{\pi}_j = \arg\max_{\pi \in \Pi} \frac{1}{|D_j|} \sum_{(x,a,r) \in D_j} 2r\mathbf{1}_{\{\pi(x)=a\}}$
  $t_{j+1} = t_j + l_j + 1$
  for $t = t_j + 1$ to $t_{j+1} - 1$ **do**
    Learner sees $X_t \in \mathcal{X}$
    Learner selects action $A_t = \hat{\pi}_j(X_t)$
    Learner receives reward $R_t = R_t^{A_t}$
  end for
end for

Theorem 2. *Suppose $R_t^a \in [0,1]$ and VC-dim$(\Pi) < \infty$. The expected regret of Algorithm 2 after J-th epoch (i.e. $T = \sum_{j=1}^{J}(1 + l_j)$) is bounded by*

$$\mathcal{R}_T \le \sum_{j=1}^{J} \left(1 + Cl_j\sqrt{\frac{\textit{VC-dim}(\Pi)}{j}}\right).$$

If choosing the epoch size $l_j = \sqrt{j/\textit{VC-dim}(\Pi)}$, then $\mathcal{R}_T$ scales as $O((\textit{VC-dim}(\Pi))^{1/3}T^{2/3})$.

Proof of Theorem 2. We first rewrite the expected regret in epoch level:

$$\begin{aligned}
\mathcal{R}_T = \mathbb{E}\Big[\sum_{t=1}^{T} R_t^{\pi^*(X_t)} - \sum_{t=1}^{T} R_t^{A_t}\Big] &= \mathbb{E}\Big[\sum_{j=1}^{J}\sum_{t=t_j}^{t_{j+1}-1}(R_t^{\pi^*(X_t)} - R_t^{A_t})\Big] \\
&= \sum_{j=1}^{J}\mathbb{E}\Big[(R_{t_j}^{\pi^*(X_{t_j})} - R_{t_j}^{A_{t_j}}) + \sum_{t=t_j+1}^{t_{j+1}-1}(R_t^{\pi^*(X_t)} - R_t^{A_t})\Big] \\
&\le \sum_{j=1}^{J} 1 + \mathbb{E}\Big[\sum_{t=t_j+1}^{t_{j+1}-1}(R_t^{\pi^*(X_t)} - R_t^{\hat{\pi}_j(X_t)})\Big]
\end{aligned}$$

Recall that we defined $t_{j+1} = t_j + l_j + 1$. Using the same trick as before, the involved expectation above can be written as

$$\begin{aligned}\mathbb{E}\Big[\sum_{t=t_j+1}^{t_{j+1}-1}(R_t^{\pi^*(X_t)} - R_t^{\hat{\pi}_j(X_t)})\Big] &= \mathbb{E}\Big[\mathbb{E}\Big[\sum_{t=t_j+1}^{t_{j+1}-1}(R_t^{\pi^*(X_t)} - R_t^{\hat{\pi}_j(X_t)})\Big|H_{t_j+1}\Big]\Big] \\ &= \mathbb{E}\left[(t_{j+1} - t_j - 1)(V(\pi^*) - V(\hat{\pi}_j))\right] \\ &= l_j\mathbb{E}\left[(V(\pi^*) - V(\hat{\pi}_j))\right]\end{aligned}$$

Note that the estimator $\hat{\pi}_j$ is built from D_j, which is of size j (one exploration per epoch). Using the same trick as in proving Theorem 1 (i.e. from inequality (1) to (3)), we have

$$\mathbb{E}\left[(V(\pi^*) - V(\hat{\pi}_j))\right] \leq C\sqrt{\frac{\text{VC-dim}(\Pi)}{j}}$$

Thus we have $\mathcal{R}_T \leq \sum_{j=1}^{J}\left(1 + Cl_j\sqrt{\frac{\text{VC-dim}(\Pi)}{j}}\right)$, as desired. Now setting $l_j = \sqrt{j/\text{VC-dim}(\Pi)}$, we have $\mathcal{R}_T \leq (1+C)J = O(J)$ and

$$T = \sum_{j=1}^{J}(l_j + 1) = J + \sum_{j=1}^{J} l_j = J + \frac{\sum_{j=1}^{J}\sqrt{j}}{\sqrt{\text{VC-dim}(\Pi)}} = J + \frac{O(J^{3/2})}{\sqrt{\text{VC-dim}(\Pi)}}$$

This implies $T \geq K\frac{J^{3/2}}{\sqrt{\text{VC-dim}(\Pi)}}$ for some universal constant K and thus $J = O((\text{VC-dim}(\Pi))^{1/3}T^{2/3})$.

□

References

[LZ08] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008.

[VDVW96] Aad W Van Der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.