STATS 710 – Seq. Decision Making with mHealth Applications Nov 1, 2016

Lecture 14: Introduction to Reinforcement Learning: Transitioning from Contextual Bandits

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Nick Seewald*

1 Review of the Contextual Bandit Problem

We begin by reviewing the contextual bandit problem in a stochastic setting. In mHealth, delayed effects of actions are often important. However, the contextual bandit problem struggles to properly account for these. Thus, we shift our attention to reinforcement learning (RL), which allows us to more easily study delayed effects, but at the cost of our ability to develop solid regret bounds.

We begin by re-introducing the contextual bandit problem using notation borrowed from RL. Consider i.i.d. context-action pairs

$$(S_j, R_j) = \left\{S_j, \left\{R_j^a\right\}_{a\in\mathcal{A}}\right\}, \quad j = 1, \ldots, M,$$

with common distribution D. Define $r_a(s) = \mathbb{E}\left[R_j^a \mid S_j = s\right]$ and $p(s) = P\left(S_j = s\right)$. A *policy* is a map $\pi : \mathcal{S} \to \mathcal{A}$ used to select actions. The action at time j is taken according to the policy π so that $A_j = \pi\left(S_j\right)$.

1.1 Setting and Optimal Policy

The general flow of information in a contextual bandit is

1: **for** $j = 1$ **to** M **do**
2: Learner sees context $S_j \in \mathcal{S}$
3: Learner selects action $A_j \in \mathcal{A}$
4: Learner receives reward $R_j = R_j^{A_j}$
5: **end for**

The problem is how to choose a policy π which selects actions A_j in order to maximize reward.

Define the space of possible policies by $\Pi = \mathcal{A}^{\mathcal{S}}$. Recall that the optimal policy π^* is that which maximizes the value function

$$V_\pi = \mathbb{E}\left[R_j^{\pi(S_j)}\right] = \mathbb{E}\left[\mathbb{E}\left(R_j^{\pi(S_j)} | S_j\right)\right] = \mathbb{E}\left[r_{\pi(S_j)}\left(S_j\right)\right].$$

Define $V_{\pi^*} = \max_{\pi\in\Pi} V_\pi$.

Proposition 1. *The optimal policy in the above setting is the one which maximizes the value function for every context $s_j \in \mathcal{S}$: $\pi^*(s) = \operatorname{argmax}_a r_a(s)$.*

Proof. We show that the value function for π^* as defined in the proposition is larger than that for

any other policy $\pi \in \Pi$.

$$\begin{aligned} V_{\pi^*} &= \mathbb{E}\left[r_{\pi^*(S_j)}(S_j)\right] \\ &= \mathbb{E}\left[\max_{a\in\mathcal{A}} r_a(S_j)\right] \\ &\geq \mathbb{E}\left[r_{\pi(S_j)}(S_j)\right] \\ &= V_\pi. \end{aligned}$$

Since $\pi^* \in \Pi$, this completes the proof. □

1.2 The Learning Algorithm

At time j, the learning algorithm determines A_j. Usually, a policy $\hat{\pi}^j$ is formed from $\mathcal{H}_{j-1}$, the history up to time j:

$$\mathcal{H}_{j-1} = \left\{S_i, A_i, R_i^{A_i} : i < j\right\} \quad \text{(and sometimes additional randomness)}.$$

We select $A_j = \hat{\pi}_{j-1}(S_j)$.

Our goal is to device a learning algorithm $\mathcal{L}$ that achieves low expected regret,

$$\begin{aligned} \mathcal{R}_M(\mathcal{L}, D, \Pi) &= \sup_{\pi\in\Pi} \mathbb{E}\left[\sum_{j=1}^M R_j^{\pi(S_j)} - \sum_{j=1}^M R_j^{A_j}\right] \\ &= \sup_{\pi\in\Pi}\left(MV_\pi - \mathbb{E}\left[\sum_{j=1}^M R_j^{A_j}\right]\right) \\ &= MV_{\pi^*} - \mathbb{E}\left[\sum_{j=1}^M R_j^{A_j}\right]. \end{aligned}$$

Using EXP3 separately for each context, we achieve $\mathcal{R}_M = O\left(\sqrt{M}\sqrt{|\mathcal{S}|\,|\mathcal{A}|\log|\mathcal{A}|}\right)$.

2 Introduction to Reinforcement Learning

Until now, the horizon for each episode $j = 1, \ldots, M$ has been 1; that is, $\left\{S_j, \left\{R_j^a\right\}_{a\in\mathcal{A}}\right\}$ are i.i.d. for $j = 1, \ldots, M$. Now, we move to the setting with M episodes (as before), but each episode has a whole series of (S_j, R_j) tuples. From now on, we now refer to contexts $S \in \mathcal{S}$ as "states". Consider i.i.d. T-tuples

$$\left\{S_{j,0}, \left\{S_{j,1}^{a_0}\right\}_{a_0\in\mathcal{A}}, \left\{S_{j,2}^{a_1}\right\}_{a_1\in\mathcal{A}}, \ldots, \left\{S_{j,T}^{a_{T-1}}\right\}_{a_{T-1}\in\mathcal{A}}\right\} \quad j = 1, \ldots, M.$$

In this setting, a policy is a vector $\pi = (\pi_0, \ldots, \pi_{T-1})^\top$ with elements $\pi_t : \mathcal{S} \to \mathcal{A}$ that is used to select actions so that

$$A_{j,0} = \pi_0(S_{j,0}),\ A_{j,1} = \pi_1\left(S_{j,1}^{A_{j,0}}\right), \ldots,\ A_{j,T-1} = \pi_{T-1}\left(S_{j,T-1}^{A_{j,T-2}}\right).$$

The distribution D of the tuples is described by the initial state distribution $p(s_0) = P\left(S_{j,0} = s_0\right)$ We also define a *transition function*

$$p_a\left(s, s'\right) = P\left(S_{j,t+1}^{A_{j,t}} = s' \mid S_{j,t}^{A_{j,t-1}} = s, A_{j,t} = a\right).$$

Note that $p_a(s, s')$ does not depend on t. This is stationarity. We assume the Markovian property; namely, that for history $\tilde{\mathcal{H}}_{j,t} = \left\{S_{j,0}, A_{j,0}, \ldots, A_{j,t-2}, S_{j,t-1}^{A_{j,t-2}}, A_{j,t-1}\right\}$,

$$P\left(S_{j,t+1}^{A_{j,t}} = s \mid \tilde{\mathcal{H}}_{j,t}, S_{j,t}^{A_{j,t-1}} = s, A_{j,t} = a\right) = p_a\left(s, s'\right).$$

Our notion of reward is a generalized version of what has been used in the past. Define $R_{j,t} = R\left(S_{j,t}^{A_{j,t-1}}, A_{j,t}, S_{j,t+1}^{A_{j,t}}\right)$, where R is some known function. In the bandit setting,

$$E\left[R_{j,t} \mid S_{j,t}^{A_{j,t-1}} = s, A_{j,t} = a\right] = \sum_{s'} R(s, a, s') p_a(s, s') := r_a(s).$$

The reward depends on the current state because the costs of taking a particular action may vary greatly with aspects of that current state. It also depends on the state following the action, since we wish to capitalize on states which may lead to actions which give higher rewards later on.

When $T = 1$ (as in the usual contextual bandit), the set of states is $\left\{S_{j,0}, \left\{S_{j,1}^{a_0}\right\}_{a_0 \in \mathcal{A}}\right\} = \{S_{j,0}, A_{j,0}, S_{j,1}\}$ and the reward received after the j^{th} episode is $R_{j,0} = R\left(S_{j,0}, A_{j,0}, S_{j,1}\right)$.

2.1 Setting and Optimal Policy

The general flow of information is

1: **for** j=1 **to** M **do**
2: Learner sees $S_{j,0} \in \mathcal{S}$
3: **for** t=0 **to** T-1 **do**
4: Learner selects action $A_{j,t} \in \mathcal{A}$
5: Learner sees state $S_{j,t+1}^{A_{j,t}} \in \mathcal{S}$
6: Learner receives reward $R_{j,t} = R\left(S_{j,t}^{A_{j,t-1}}, A_{j,t}, S_{j,t+1}^{A_{j,t}}\right)$
7: **end for**
8: **end for**

Consider the policy space $\Pi = \left(\mathcal{A}^{\mathcal{S}}\right)^T$. We redefine our notion of the value function, which is commonly conditioned on the initial state:

$$V_\pi^T\left(s_0\right) = \mathbb{E}_{A \sim \pi}\left[\sum_{t=0}^{T-1} R_{j,t} \mid S_{j,0} = s_0\right]. \tag{1}$$

The optimal policy π^* should maximize $V_\pi^T(s_0)$.

Proposition 2. *The optimal policy $\pi^*(s)$ is that which, for every time $t = 1, \ldots, T$, satisfies $\pi_{T-t}^*(s) = \operatorname{argmax}_a r_a(s) + \sum_{s'} p_a(s, s') V_{\pi_{T-t+1}^*}^{t-1}(s')$, where $V_{\pi_{T-t}^*}^t(s)$ is defined to be $\max_a r_a(s) + \sum_{s'} p_a(s, s') V_{\pi_{T-t+1}^*}^{t-1}(s')$ and $V_\pi^0(s) := 0$. Everything is defined recursively from t= 1 to t = T.*

Proof. We begin with the definition of $\max_\pi V_\pi^T(s_0)$, even though we do not write it, below all expectations are conditional on $S_{j,0} = s_0$:

$$\begin{aligned}\max_\pi V_\pi^T(s_0) &= \max_{\pi_0}\max_{\pi_1}\cdots\max_{\pi_{T-1}}\mathbb{E}\left[\sum_{t=0}^{T-1} R_{j,t} \mid S_{j,0}=s_0\right] \\ &= \max_{\pi_0}\cdots\max_{\pi_{T-1}}\mathbb{E}_{A\sim\pi}\left[\sum_{t=0}^{T-2} R_{j,t} + \mathbb{E}\left[R_{j,T-1} \mid S_{j,T-1}, A_{j,T-1}=\pi_{T-1}(S_{j,T-1})\right]\right] \quad (2)\end{aligned}$$

Notice that the internal expectation in eq. (2) can be rewritten as $r_{\pi_{T-1}(S_{j,T-1})}(S_{j,T-1})$, and select $\pi^*_{T-1}(s) = \text{argmax}_a\, r_a(s)$. Then we have

$$\max_\pi V_\pi^T(s_0) \le \max_{\pi_0}\cdots\max_{\pi_{T-2}}\mathbb{E}_{A\sim\pi}\left[\sum_{t=0}^{T-2} R_{j,t} + r_{\pi^*_{T-1}(S_{j,T-1})}(S_{j,T-1})\right] \quad (3)$$

Define $V^1_{\pi_{T-1}}(s) = r_{\pi_{T-1}(s)}(s)$, so that $V^1_{\pi^*_{T-1}}(s) = r_{\pi^*_{T-1}(s)}(s) = \max_a r_a(s)$. Now the right-hand side of eq. (3), denoted RHS(3), is

$$\begin{aligned}\text{RHS(3)} &= \max_{\pi_0}\cdots\max_{\pi_{T-2}}\mathbb{E}_{A\sim\pi}\left[\sum_{t=0}^{T-2} R_{j,t} + V^1_{\pi^*_{T-1}}(S_{j,T-1})\right] \\ &\le \max_{\pi_0}\cdots\max_{\pi_{T-2}}\mathbb{E}_{A\sim\pi}\left[\sum_{t=0}^{T-3} R_{j,t} + \mathbb{E}\left[R_{j,T-2} + V^1_{\pi^*_{T-1}}(S_{j,T-1}) \mid S_{j,T-2}, A_{j,T-2}=\pi_{T-2}(S_{j,T-2})\right]\right]. \quad (4)\end{aligned}$$

As above, notice that the internal expectation in eq. (4) is equal to

$$r_{\pi_{T-2}(S_{j,T-2})}(S_{j,T-2}) + \sum_{s'} p_{\pi_{T-2}(S_{j,T-2})}(S_{j,T-2}, s')\, V^1_{\pi^*_{T-1}}(s').$$

Select $\pi^*_{T-2}(s) = \text{argmax}_a\left(r_a(s) + \sum_{s'} p_a(s,s')\, V^1_{\pi^*_{T-1}}(s')\right)$. Then

$$\text{RHS(4)} \le \max_{\pi_0}\cdots\max_{\pi_{T-3}}\mathbb{E}_{A\sim\pi}\left[\sum_{t=0}^{T-3} R_{j,t} + V^2_{\pi^*_{T-2}}(S_{j,T-2})\right],$$

where $V^2_{\pi^*_{T-2}}(s) = \max_a\left(r_a(s) + \sum_{s'} p_a(s,s') V^1_{\pi^*_{T-1}}(s')\right)$. Continuing in this way, by iteratively defining

$$\begin{aligned}\pi^*_{T-t}(s) &:= \operatorname*{argmax}_a r_a(s) + \sum_{s'} p_a(s,s') V^{t-1}_{\pi^*_{T-t+1}}(s') \\ V^t_{\pi^*_{T-t}}(s) &:= \max_a r_a(s) + \sum_{s'} p_a(s,s') V^{t-1}_{\pi^*_{T-t+1}}(s')\end{aligned}$$

for $t = 3, \ldots, T$ we arrive at

$$\max_\pi V_\pi^T(s_0) \le V^T_{\pi^*}(s_0)$$

Since $\pi^* \in \Pi$ by definition, we must have that it is the optimal policy. □