Algorithms and Uncertainty, Winter 2018/19 Lecture 7 (4 pages)

Optimal Stopping

Instructor: Thomas Kesselheim

Let us consider the following *online selection problem*, in which you have to make commitments before you know all you choices. Suppose you want to buy a house. You go see several houses and (a bit simplifying here) after each visit you have to decide immediately and irrevocably if you want to buy this particular house or if you want to keep on looking – then somebody else will buy it. Another motivation would be that you want to find the love of your life. You start dating and (even more simplifying here) after each first date you have to decide whether you you want to marry this person or if you want to keep looking.

We can model this problem as follows. There are n candidates of values $v_1, \dots v_n \in \mathbb{R}$, $v_i \geq 0$. You see the values of these candidates in order $1, \dots, n$. After having seen the i-th candidate you can choose to select it or to reject it. The goal is to maximize the value of the candidate that you select.

1 Known Distributions

You might have heard of this problem as the *secretary problem*. Here, the candidates arrive in random order and you have no prior knowledge. We make a different assumption today: We assume that each v_i is drawn from a probability distribution. The probability distributions may differ between different rounds. All distributions are known in advance. Each draw is independent.

Example 7.1. *The values v_i could come from the following distributions.*

$$v_1 \sim \text{Uniform}\{0, 1, 2, 10\} \qquad v_2 \sim \text{Uniform}\{0, 3\}$$
$$v_3 \sim \text{Uniform}\{0, 1, 2, 3\} \qquad v_4 \sim \text{Uniform}\{1, 2\}$$

We would definitely accept a 10 in the first step. But would we accept a value of 2? Or would we take our chance to get a 3 later on? After all, $\mathbf{E}[v_2] = \mathbf{E}[v_3] = \mathbf{E}[v_4] = \frac{3}{2}$.

This problem can be modeled as a Markov decision process. There are two actions STOP and CONTINUE. The states are a little more complicated. In the state space, we have to store the current i and v_i that we are deciding on. If we choose action STOP in state (i, v_i), we get a reward of v_i and move to state STOPPED. If we choose action CONTINUE, then we get no reward and move to state $(i+1, v_{i+1})$. The state STOPPED simply means that we have already stopped the sequence, so any action makes us remain in the state and gives no reward. To get the initialization correct, we start from state $(0, 0)$ and have a time horizon of $T = n + 1$. To avoid technicalities, we assume that each v_i can attain only finitely many values. Then the state space can also be finite, namely $(\{0, 1, \dots, n\} \times X) \cup \{\text{STOPPED}\}$, where X is the set of possible values for $v_1, \dots, v_n$. In principle, all of our results today also hold for $X = \mathbb{R}_{\geq 0}$.

2 Characterizing the Optimal Policy

We will now characterize the optimal policies. Recall that we defined $V^*(s, T)$ to be the expected reward of an optimal policy started at state s and running for T steps. We derived that

$$V^*(s, T) = \max_{a \in \mathcal{A}} \left(r_a(s) + \sum_{s' \in \mathcal{S}} p_a(s, s') V^*(s', T-1) \right) \quad . \tag{1}$$

In our particular case, we are in states (i, v_i) when there are $n-i+1$ more steps to go unless we have already stopped. There are only two actions: One gives immediate reward v_i and none in the future; the other one only moves on to $(i+1, v_{i+1})$. So, if in (i, v_i), the optimal policy chooses STOP, then

$$v_i \geq \sum_y \mathbf{Pr}\left[v_{i+1} = y\right] V^*((i+1, y), T-i-1) \ ,$$

if it chooses CONTINUE, then

$$v_i \leq \sum_y \mathbf{Pr}\left[v_{i+1} = y\right] V^*((i+1, y), T-i-1) \ ,$$

Note that the right-hand side is nothing but the expected reward of an optimal policy that sees only the subinstance $v_{i+1}, \ldots, v_n$. In particular, the right-hand side is independent of v_i. This means that we simply have a threshold for v_i, which it has to exceed in order to be accepted.

Theorem 7.2. *It is an optimal policy to use thresholds $\tau_1 \leq \ldots \leq \tau_n$ such that in the i-th step we accept if $v_i > \tau_i$ and reject if $v_i < \tau_i$. The thresholds are defined recursively by $\tau_n = 0$ and $\tau_i = \mathbf{E}\left[\max\{v_{i+1}, \tau_{i+1}\}\right]$ for $i < n$.*

The optimal policy is unique except for the tie-breaking at $\tau_i = v_i$, which is irrelevant.

Proof. We will show that the sequence τ_i as defined in the theorem is exactly the expected reward of an optimal policy on only $v_{i+1}, \ldots, v_n$. We show this by induction on i downward from n to 1. The induction base $i = n$ is trivial because the policy's reward is 0 on an empty sequence.

For the induction step, we assume that τ_i is the expected reward of an optimal policy on $v_{i+1}, \ldots, v_n$ and we would live to derive the respective statement for $i-1$. Recall that, by Equation (1), an optimal policy's expected reward on $v_i, \ldots, v_n$ for a fixed v_i is

$$V^*((i, v_i), n-i+1) = \max_{a \in \mathcal{A}} \left(r_a(s) + \sum_{s' \in \mathcal{S}} p_a((i, v_i), s') V^*(s', n-i) \right) \ ,$$

for $a = \text{STOP}$, the term is exactly v_i; for $a = \text{CONTINUE}$, it is $\mathbf{E}\left[V^*((i+1, v_{i+1}), n-i)\right]$. By induction hypothesis $\mathbf{E}\left[V^*((i+1, v_{i+1}), n-i)\right] = \tau_i$ by induction hypothesis. Therefore,

$$V^*((i, v_i), n-i+1) = \max\{v_i, \tau_i\}$$

for all v_i. Taking the expectation over v_i, this completes the induction.

We can also derive the structure of the policy. If $v_i > \tau_i$, the maximum is attained at $a = \text{STOP}$. So this is what the optimal policy has to do. If $v_i < \tau_i$, the maximum is attained at $a = \text{CONTINUE}$. If $v_i = \tau_i$, both actions give the same expected reward, so any choice leads to an optimal policy.

Note that also the sequence of thresholds is non-increasing. This follows immediately from the recursion. There is also another, intuitive explanation: We set τ_i to be the reward of an optimal policy on only $v_{i+1}, \ldots, v_n$. One such policy would be to ignore v_{i+1} and only operate on $v_{i+2}, \ldots, v_n$. Therefore $\tau_i \geq \tau_{i+1}$. □

Example 7.3. *We derive an optimal policy for the example above with*

$$v_1 \sim \text{Uniform}\{0, 1, 2, 10\} \qquad v_2 \sim \text{Uniform}\{0, 3\}$$
$$v_3 \sim \text{Uniform}\{0, 1, 2, 3\} \qquad v_4 \sim \text{Uniform}\{1, 2\}$$

We get $\tau_4 = 0$ *and* $\tau_3 = \mathbf{E}[\max\{v_4, 0\}] = \mathbf{E}[v_4] = \frac{3}{2}$.
Deriving τ_2 *is a little more complicated. It is*

$$\tau_2 = \mathbf{E}\left[\max\left\{v_3, \frac{3}{2}\right\}\right] = \frac{1}{2}\cdot\frac{3}{2} + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 3 = 2 \ .$$

For τ_1*, we therefore get*

$$\tau_1 = \mathbf{E}[\max\{v_2, 2\}] = \frac{1}{2}\cdot 2 + \frac{1}{2}\cdot 3 = \frac{5}{2} \ .$$

We can also derive the expected reward of this optimal policy as

$$\mathbf{E}[\max\{v_1, \tau_1\}] = \frac{3}{4}\cdot\frac{5}{2} + \frac{1}{4}\cdot 10 = \frac{95}{8} \ .$$

3 Comparison to the Offline Optimum

Clearly, one could get a higher reward than the optimal policy if one knew the entire sequence $v_1, \ldots, v_n$ in advance. Then, one would stop the sequence at its highest value and get reward $\max_i v_i$. This is exactly the definition of the offline optimum in competitive analysis. But how much better can the offline optimum be? Interestingly, the factor is bounded here.

Theorem 7.4. *The optimal policy's expected reward is at least* $\frac{1}{2}\mathbf{E}[\max_i v_i]$.

Such a theorem is called a "prophet inequality". The result was developed independent of competitive analysis. When talking about the offline optimum, the researchers had a prophet in mind, who knows the future and make decisions depending on this. Any (online) policy is only a gambler, who does not know the future but can take a chance on the outcome.

There is a very elegant and simple proof of this prophet inequality. Rather than establishing the bound for the (complicated) optimal policy, we will define a much simpler policy, which we can talk about more easily.

The simple policy is defined as follows. Set a threshold $\tau = \frac{1}{2}\mathbf{E}[\max_i v_i]$. Stop the sequence the first time we see a value $v_i \geq \tau$.

The simple policy is clearly suboptimal. For example, even in the last step the threshold applies. So, if we reach the last step and $v_n < \tau$, then the policy gets no reward, although it could easily collect v_n. Nonetheless, the guarantee holds also for this policy and the optimal policy's reward can only be higher.

Proposition 7.5. *The single-threshold policy's expected reward* $V(\pi)$ *is at least* $\frac{1}{2}\mathbf{E}[\max_i v_i]$.

Proof. Let q be the probability that the sequence is not stopped at all. That is, $q = \mathbf{Pr}[v_1 < \tau, \ldots, v_n < \tau]$. Furthermore, for all i, define a random variable u_i by setting $u_i = v_i - \tau$ if the sequence is stopped at i and $u_i = 0$ otherwise. The expected reward of the policy is given by

$$V(\pi) = \mathbf{E}\left[\sum_i u_i\right] + (1-q)\tau \ . \tag{2}$$

This is because at most one of the u_i will be positive and its value will be by how much the threshold was exceeded. We have to add the threshold but only with the probability that at least one number beats it.

Depending on your background, the following interpretation may be helpful. We sell one item at a price of τ among buyers of values $v_1, \ldots, v_n$. The first buyer willing to pay τ gets the

item and pays τ. Then u_i will be the respective buyer's utility and $(1-q)\tau$ will be the expected revenue.

For all i, we have $u_i = \max\{v_i - \tau, 0\} \cdot \mathbf{1}_{v_1<\tau,\dots,v_{i-1}<\tau}$, where $\mathbf{1}$ denotes the 0/1 indicator. Therefore, by independence, we have

$$\begin{aligned}\mathbf{E}[u_i] &= \mathbf{E}\left[\max\{v_i - \tau, 0\} \cdot \mathbf{1}_{v_1<\tau,\dots,v_{i-1}<\tau}\right] \\ &= \mathbf{E}[\max\{v_i - \tau, 0\}]\,\mathbf{Pr}[v_1 < \tau, \dots, v_{i-1} < \tau] \ .\end{aligned}$$

Note that $\mathbf{Pr}[v_1 < \tau, \dots, v_{i-1} < \tau] \geq \mathbf{Pr}[v_1 < \tau, \dots, v_n < \tau] = q$, and so

$$\mathbf{E}[u_i] \geq \mathbf{E}[\max\{v_i - \tau, 0\}]\, q \ .$$

Taking the sum over all i, linearity of expectation gives us

$$\mathbf{E}\left[\sum_i u_i\right] \geq \mathbf{E}\left[\sum_i \max\{v_i - \tau, 0\}\right] q \ .$$

Furthermore, every sum of non-negative numbers is at least the maximum of these numbers, which means

$$\sum_i \max\{v_i - \tau, 0\} \geq \max_i \max\{v_i - \tau, 0\} \geq \max_i v_i - \tau \ .$$

This gives us

$$\mathbf{E}\left[\sum_i u_i\right] \geq \mathbf{E}\left[\max_i v_i - \tau\right] q = \left(\mathbf{E}\left[\max_i v_i\right] - \tau\right) q \ .$$

Plugging this into Equation (2), we get

$$V(\pi) \geq \left(\mathbf{E}\left[\max_i v_i\right] - \tau\right) q + (1-q)\tau \ .$$

This statement holds for all choices of τ. Using $\tau = \frac{1}{2}\mathbf{E}[\max_i v_i]$, we get

$$V(\pi) \geq \frac{1}{2}\mathbf{E}\left[\max_i v_i\right] q + (1-q)\frac{1}{2}\mathbf{E}\left[\max_i v_i\right] = \frac{1}{2}\mathbf{E}\left[\max_i v_i\right] \ ,$$

regardless of q. Therefore, the claim holds. □

This guarantee is optimal.

Proposition 7.6. *The guarantee in Theorem 7.4 cannot be improved.*

Proof. Consider $v_1 = 1$ with probability 1, $v_2 = \frac{1}{\epsilon}$ with probability ϵ, $v_2 = 0$ otherwise. There are effectively two policies: Stop at $v_1 = 1$ or continue. Both policies have an expected reward of 1 but $\mathbf{E}[\max_i v_i] = (1-\epsilon) + \epsilon\frac{1}{\epsilon} = 2 - \epsilon$. This holds for all $\epsilon > 0$, so the ratio gets arbitrarily close to 1. □