Algorithms and Uncertainty, Winter 2025/26 Lecture 16 (4 pages)

Learning for Pandora's Box

Thomas Kesselheim Last Update: December 7, 2025

In our discussion of Markov decision processes, we were assuming that probability distributions are known beforehand. But why is this a reasonable assumption? Most likely, we can use observations from the past. But how and why can we use past observations? And how much knowledge do we need to have? Today, we will discuss these questions considering Pandora's Box.

1 Recap: Pandora's Box (with Known Distributions)

We have n boxes. Box i contains a prize of value v_i. We don't know v_i but only its distribution until we open the box. Opening box i costs c_i. The final reward is given as

$$\max_{i:\text{box } i \text{ opened}} v_i - \sum_{i:\text{box } i \text{ opened}} c_i \ ,$$

where we define the maximum as 0 if no boxes are opened.

We showed that the fair-cap policy is optimal. For box i, let the fair cap σ_i be defined by $\mathbf{E}\left[\max\{0, v_i - \sigma_i\}\right] = c_i$. (We called $\max\{0, v_i - \sigma_i\}$ the bonus of box i.) The policy opens the boxes by decreasing fair cap σ_i. It stops when the largest observed value v_{i^*} exceeds the highest remaining cap and selects i^*.

We let V^* denote the expected reward of the optimal policy.

2 Incorrect Costs

Before we come to the problem of learning the optimal policy, we first show a useful lemma. It helps us understand the following setting: Suppose the true costs are $c_1, \dots, c_n$ but you run the policy which would be optimal for costs $c'_1, \dots, c'_n$. What effect does this have?

Lemma 16.1. *Consider a policy π' that would be optimal if costs were $c'_1, \dots, c'_n$. If $|c_i - c'_i| \leq \gamma$ for all boxes i, then the expected reward of π' with respect to actual costs $c_1, \dots, c_n$ is at least $V^* - 2n\gamma$.*

Proof. For any policy π let $V(\pi, c)$ and $V(\pi, c')$ denote the expected rewards when the cost are c or c', respectively.

We now have

$$\begin{aligned}
V(\pi, c) &= \mathbf{E}\left[\max_{i:\pi \text{ opens box } i} v_i - \sum_{i:\pi \text{ opens box } i} c_i\right] \\
&\geq \mathbf{E}\left[\max_{i:\pi \text{ opens box } i} v_i - \sum_{i:\pi \text{ opens box } i} (c'_i + \gamma)\right] \\
&\geq \mathbf{E}\left[\max_{i:\pi \text{ opens box } i} v_i - \sum_{i:\pi \text{ opens box } i} c'_i\right] - n\gamma \\
&= V(\pi, c') - n\gamma
\end{aligned}$$

and analogously $V(\pi, c) \leq V(\pi, c') + n\gamma$.

Let π^* be the optimal policy for cost c, π' be the optimal one for costs c'. Then we have

$$V(\pi', c) \geq V(\pi', c') - n\gamma \geq V(\pi^*, c') - n\gamma \geq V(\pi^*, c) - 2n\gamma \ .$$

□

3 Learning from Samples

Let's formalize our main question for today as follows. For every box i, we are given T samples $v_i^{(1)}, \ldots, v_i^{(T)}$ from the prize distribution. Based on these samples, we would like to find a good policy.

Indeed, there is a very natural way to choose a policy: The T samples define another probability distribution, called the *empirical distribution*: Simply draw one of $v_i^{(1)}, \ldots, v_i^{(T)}$ uniformly at random. Let $\tilde{\sigma}_i$ be the fair cap of box i with respect to the empirical distribution. That is, $\tilde{\sigma}_i$ is chosen such that

$$\frac{1}{T}\sum_{t=1}^{T} \max\{0, v_i^{(t)} - \tilde{\sigma}_i\} = c_i \ .$$

Our learned policy now proceeds like the fair-cap policy on $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_n$. That is, it opens the boxes by decreasing empirical fair cap $\tilde{\sigma}_i$. It stops when the largest observed value v_{i^*} exceeds the highest remaining cap and selects i^*.

In order to analyze the expected reward of this policy, we will assume that the prizes are upper-bounded by 1; so, $v_i \in [0, 1]$ with probability 1 for all i. It is straightforward to apply the result to settings where $v_i \in [0, \rho]$ with probability 1 for a $\rho > 0$ by scaling.

Theorem 16.2. *Let $v_i \in [0, 1]$ with probability 1 for all i. For all $\epsilon, \delta > 0$, if $T \geq \frac{2n^2 \ln(2n/\delta)}{\epsilon^2}$, then the expected reward of the learned policy is at least $V^* - \epsilon$ with probability at least $1 - \delta$.*

To prove this theorem, we will use two standard inequalities without proofs. The union bound gives an easy upper bound on the probability of a union of events.

Lemma 16.3 (Union Bound)**.** *For any sequence of not necessarily disjoint events $\mathcal{E}_1, \mathcal{E}_2 \ldots$, we have*

$$\mathbf{Pr}\left[\mathcal{E}_1 \cup \mathcal{E}_2 \cup \ldots\right] \leq \mathbf{Pr}\left[\mathcal{E}_1\right] + \mathbf{Pr}\left[\mathcal{E}_2\right] + \ldots \ .$$

Hoeffding's inequality is a quantitative version of the law of large numbers. It states that we get close to the expectation if we take the average of sufficiently many independent draws from a distribution.

Lemma 16.4 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_N$ be independent random variables such that $a_i \leq Z_i \leq b_i$ with probability 1. Let $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ be their average. Then for all $\gamma \geq 0$*

$$\mathbf{Pr}\left[\bar{Z} - \mathbf{E}\left[\bar{Z}\right] \geq \gamma\right] \leq \exp\left(-\frac{2N^2\gamma^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \ ,$$

$$\mathbf{Pr}\left[\bar{Z} - \mathbf{E}\left[\bar{Z}\right] \leq -\gamma\right] \leq \exp\left(-\frac{2N^2\gamma^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \ ,$$

and

$$\mathbf{Pr}\left[|\bar{Z} - \mathbf{E}\left[\bar{Z}\right]| \geq \gamma\right] \leq 2\exp\left(-\frac{2N^2\gamma^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \ .$$

As a side remark, note that the second inequality follows by applying the first one on $-Z_1, \dots, -Z_N$. The third one follows by combining the two using the Union Bound.

4 Learning a Single Fair Cap

We will now prove our key lemma, which is only for a single box. We will show that if T is large enough then it is very likely that $\mathbf{E}[\max\{0, v_i - \tilde{\sigma}_i\}]$ and c_i are close. Note that it is not immediately clear how to draw such a conclusion from Hoeffding's inequality because $\tilde{\sigma}_i$ is the solution to the inequality $\frac{1}{T}\sum_{t=1}^{T} \max\{0, v_i^{(t)} - \tilde{\sigma}_i\} = c_i$.

Lemma 16.5. *For every box i, with probability at least $1 - 2\exp(-T\gamma^2)$, the cap $\tilde{\sigma}_i$ determined from the samples fulfills*

$$|\mathbf{E}[\max\{0, v_i - \tilde{\sigma}_i\}] - c_i| \leq \gamma \ ,$$

where the expectation is taken over v_i.

Proof. Consider the function $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $f(x) = \mathbf{E}[\max\{0, v_i - x\}]$. We will show that with probability at most $2\exp(-T\gamma^2)$, we have

$$f(\tilde{\sigma}_i) > c_i + \gamma \quad \text{or} \quad f(\tilde{\sigma}_i) < c_i - \gamma \ ,$$

where $\tilde{\sigma}_i$ is a random variable that depends on our samples.

Let's first bound the probability that $f(\tilde{\sigma}_i) > c_i + \gamma$. The function f is continuous and non-increasing. Therefore, the event can only take place if $f(0) > c_i + \gamma$. In this case, there is an $a > 0$ such that $f(a) = c_i + \gamma$ and $f(\tilde{\sigma}_i) > c_i + \gamma$ is equivalent to $\tilde{\sigma}_i < a$.

Let's define $\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} \max\{0, v_i^{(t)} - x\}$. Also this function is continuous and non-decreasing. Recall that we choose $\tilde{\sigma}_i$ such that $\hat{f}(\tilde{\sigma}_i) = c_i$. By monotonicity, if $\tilde{\sigma}_i < a$ then $\hat{f}(a) \leq \hat{f}(\tilde{\sigma}_i) = c_i$.

So, we will bound the probability that $\hat{f}(a) \leq c_i$. To this end, we use Hoeffding's inequality. Let $Z_t = \max\{0, v_i^{(t)} - a\}$. Note that $0 \leq Z_t \leq 1$ and that $\mathbf{E}[Z_t] = \mathbf{E}\left[\max\{0, v_i^{(t)} - a\}\right] = f(a) = c_i + \gamma$. Therefore, Hoeffding's inequality tells us for $\hat{f}(a) = \bar{Z} = \frac{1}{T}\sum_{t=1}^{T} Z_t$ that

$$\mathbf{Pr}\left[\hat{f}(a) \leq c_i\right] = \mathbf{Pr}\left[\bar{Z} \leq \mathbf{E}\left[\bar{Z}\right] - \gamma\right] \leq \exp(-2T\gamma^2) \ .$$

In combination, we have (even when $f(0) \leq c_i + \gamma$)

$$\mathbf{Pr}[f(\tilde{\sigma}_i) > c_i + \gamma] \leq \exp(-2T\gamma^2) \ .$$

Analogously, to bound the probability that $f(\tilde{\sigma}_i) < c_i - \gamma$, we distinguish whether $\lim_{x\to\infty} f(x) < c_i - \gamma$ or not. Only in the former case it can happen that $f(\tilde{\sigma}_i) < c_i - \gamma$. In this case, there is a $b \geq 0$ such that $f(b) = c_i - \gamma$. Hoeffding's inequality now gives us

$$\mathbf{Pr}[f(\tilde{\sigma}_i) < c_i - \gamma] \leq \mathbf{Pr}\left[\hat{f}(b) \geq c_i\right] \leq \exp(-2T\gamma^2) \ .$$

Applying a union bound, we get

$$\mathbf{Pr}[f(\tilde{\sigma}_i) > c_i + \gamma \text{ or } f(\tilde{\sigma}_i) < c_i - \gamma] \leq 2\exp(-2T\gamma^2) \ .$$

This proves the claim. □

5 Putting the Pieces Together

We can now complete the proof of Theorem 16.2. The idea is that the policy computed based on samples can be understood as an optimal policy with respect to costs that are slightly off.

Proof of Theorem 16.2. Let $\gamma = \frac{\epsilon}{2n}$, $T \geq \frac{2n^2 \ln(2n/\delta)}{\epsilon^2} = \frac{\ln(2n/\delta)}{2\gamma^2}$. By union bound, we have

$$\begin{aligned}
&\mathbf{Pr}\left[\exists i : |\mathbf{E}\left[\max\{0, v_i - \tilde{\sigma}_i\}\right] - c_i| > \gamma\right] \\
&\leq \sum_{i=1}^{n} \mathbf{Pr}\left[|\mathbf{E}\left[\max\{0, v_i - \tilde{\sigma}_i\}\right] - c_i| > \gamma\right] \\
&\leq n2\exp(-2T\gamma^2) \\
&\leq \delta
\end{aligned}$$

So, with probability at least $1 - \delta$, we have for all boxes i

$$|\mathbf{E}\left[\max\{0, v_i - \tilde{\sigma}_i\}\right] - c_i| \leq \gamma \ .$$

Consider any fixed choice of $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_n$ such that this happens. Let $c'_i = \mathbf{E}\left[\max\{0, v_i - \tilde{\sigma}_i\}\right]$. We have $|c'_i - c_i| \leq \gamma$ for all i.

Now $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_n$ correspond to an optimal policy with respect to costs $c'_1, \ldots, c'_n$. By Lemma 16.1, we know that its expected reward is $V^* - 2n\gamma = V^* - \epsilon$. □

6 Can we do better?

We have seen that $O(\frac{n^2 \log n}{\epsilon^2})$ samples suffices to get within ϵ of the optimal policy. This raises the question if we can do better. Indeed, if one is a little more careful and also uses stronger concentration bounds, the dependence on n can be made $n \log n$ instead of $n^2 \log n$.

However, the dependence of ϵ cannot be improved. The counterexample is surprisingly simple. Suppose you have only a single box, which always costs $c_1 = \frac{1}{2}$ to open. There are two possible distributions: Either $v_1 = 1$ with probability $\frac{1}{2} + \epsilon$ and 0 otherwise or $v_1 = 1$ with probability $\frac{1}{2} - \epsilon$ and 0. In case of the former distribution, the optimal policy opens the box; in case of the latter, it does not. The respective other policy's reward is ϵ worse. However, with $o(\frac{1}{\epsilon^2})$ samples, one cannot distinguish the two.

Another question is: What if we get less information, namely we only get samples from boxes that we opened in the past. This is an question of online learning, where we get a tradeoff between exploration and exploitation.

References and Further Reading

- The proof presented here is from joint work with Sahil Singla but not published.
- Chenghao Guo, Zhiyi Huang, Zhihao Gavin Tang, Xinzhi Zhang: Generalizing Complex Hypotheses on Product Distributions: Auctions, Prophet Inequalities, and Pandora's Problem. COLT 2021: 2248-2288 (Gives an optimal bound for Pandora's Box but with a different approach.)
- Khashayar Gatmiry, Thomas Kesselheim, Sahil Singla, Yifan Wang: Bandit Algorithms for Prophet Inequality and Pandora's Box. SODA 2024 (A generalized version of the approach taken here, which can deal with limited feedback.)