Machine Learning Theory (CSC 431/531) - Lectures 21 and 22

Nishant Mehta

1 The non-stochastic multi-armed bandit problem

Thus far, we have looked at online learning games of full information. In a full information setting, at the end of each round Learner gets to observe the loss of any action it could have taken. We now switch to games of partial information. Many such games have been studied, and the specific one we will cover is the class of multi-armed bandit problems. In the non-stochastic multi-armed bandit problem, also called the adversarial multi-armed bandit problem, in each of a sequence of rounds:

1. Learner pulls one arm I_t from a set of K arms, possibly by randomizing according to a distribution $p_t \in \Delta_K$.
2. Nature plays a loss vector $\ell_t \in [0,1]^K$ which assigns loss $\ell_{j,t}$ to each arm $j \in [K]$.
3. Learner suffers loss $\ell_{I_t,t}$ and observes only this loss.

If Nature is an adaptive adversary, then it can play ℓ_t with knowledge of Learner's realizations $I_1, I_2, \ldots, I_{t-1}$ as well as the distribution p_t.

The above problem is equivalent to decision-theoretic online learning if *(a)* Learner ultimately must play a single expert in each round and *(b)* Learner's feedback is limited to *bandit feedback*, i.e., Learner only observes the loss of the expert it plays.

Using the above notation, Learner's regret is

$$R_T = \sum_{t=1}^{T} \ell_{I_t,t} - \min_{j \in [K]} \sum_{t=1}^{T} \ell_{j,t}.$$

It turns out that any learning algorithm that obtains low regret must necessarily randomize, even if Nature is only oblivious, and so for any successful algorithm for this setting, I_t will be a random variable. To see how a deterministic strategy can fail to obtain low regret, we consider a simple example.

1.1 The need for randomization

Let $K = 2$, and suppose that the learning algorithm is deterministic, so that conditional on $\ell_1, \ldots, \ell_{t-1}$, the learning algorithm always plays a fixed action I_t. Then in round t, Nature sets the loss vector as follows:

$$\ell_t = \begin{cases} (1,0) & \text{if } I_t = 1 \\ (0,1) & \text{if } I_t = 2. \end{cases} \tag{1}$$

Then, on the one hand, we have

$$\sum_{t=1}^{T} \ell_{I_t,t} = T,$$

while on the other hand, we have

$$\sum_{t=1}^{T}\sum_{j=1}^{2} \ell_{j,t} = T \quad \Longrightarrow \quad \min_{j=1,2} \sum_{t=1}^{T} \ell_{j,t} \leq \frac{T}{2},$$

and so the regret exhibits the hopelessly linear growth $\frac{T}{2}$.

Moreover, Nature is oblivious since it can simulate the deterministic learning algorithm to identify a sequence of losses satisfying (1) for all $t \in [T]$.

1.2 Expected regret and pseudo-regret

Because the learning algorithm must (and Nature may) randomize, our interest will be in studying regret bounds that hold in expectation. It is also possible to develop bounds that hold with high probability, and such bounds are important to have in practice; for simplicity, we forego an analysis that gives high probability guarantees.

The *expected regret* is

$$\mathsf{E}[R_T] = \mathsf{E}\left[\sum_{t=1}^{T} \ell_{I_t,t} - \min_{j\in[K]} \sum_{t=1}^{T} \ell_{j,t}\right].$$

A related notion of regret is known as the *pseudo-regret*, defined as

$$\bar{R}_T = \mathsf{E}\left[\sum_{t=1}^{T} \ell_{I_t,t}\right] - \min_{j\in[K]} \mathsf{E}\left[\sum_{t=1}^{T} \ell_{j,t}\right].$$

In this first study of the non-stochastic setting, our focus will be on obtaining bounds on the pseudo-regret rather than the expected regret, because:

1. It is simpler to upper bound the pseudo-regret;
2. If Nature is oblivious, an upper bound on the worst-case pseudo-regret is also an upper bound on the worst-case expected regret.

The first observation is true because

$$\bar{R}_T \leq \mathsf{E}[R_T], \tag{2}$$

which follows from the rewrite

$$\max_{j\in[K]} \mathsf{E}\left[\sum_{t=1}^{T} \ell_{I_t,t} - \sum_{t=1}^{T} \ell_{j,t}\right] \leq \mathsf{E}\left[\max_{j\in[K]} \left\{\sum_{t=1}^{T} \ell_{I_t,t} - \sum_{t=1}^{T} \ell_{j,t}\right\}\right], \tag{3}$$

An upper bound on the expected regret is thus also an upper bound on the pseudo-regret.

Let us see why the second observation is true. First, suppose that Nature is deterministic (and oblivious); this is a special case of an oblivious adversary. The pseudo-regret then reduces to

$$\bar{R}_T = \mathsf{E}\left[\sum_{t=1}^{T} \ell_{I_t,t}\right] - \min_{j\in[K]} \sum_{t=1}^{T} \ell_{j,t} = \mathsf{E}\left[\sum_{t=1}^{T} \ell_{I_t,t} - \min_{j\in[K]} \sum_{t=1}^{T} \ell_{j,t}\right] = \mathsf{E}[R_T], \tag{4}$$

which is just the expected regret. Thus, in the special case of deterministic (oblivious) adversaries, the pseudo-regret is equal to the expected regret. Next, suppose that Nature is oblivious but might also randomize. Let $\mathbf{B}$ denote the randomization of Nature. Then, since Nature is oblivious,

$$\mathsf{E}\left[R_T\right] = \mathsf{E}_{\mathbf{B}}\left[\mathsf{E}\left[R_T \mid \mathbf{B}\right]\right].$$

Next, observe that since the learning algorithm is fixed, it holds that

$$\mathsf{E}_{\mathbf{B}}\left[\mathsf{E}\left[R_T \mid \mathbf{B}\right]\right] \leq \sup_{\ell_1,\ldots,\ell_T} \mathsf{E}\left[R_T\right];$$

we thus see that the expected reget under an oblivious adversary is upper bounded by the worst-case expected regret under a deterministic (oblivious) adversary; also, the inequality becomes an equality if we instead consider the worst-case expected regret under an oblivious adversary.

Combining this fact with (4), we have

$$\sup_{\text{oblivious}} \mathsf{E}\left[R_T\right] = \sup_{\text{deterministic}} \mathsf{E}\left[R_T\right] = \sup_{\text{deterministic}} \bar{R}_T; \tag{5}$$

the first supremum is taken over all oblivious adversaries, while the other supremums are taken over all deterministic oblivious adversaries.

Thus, in order to upper bound the worst-case expected regret under a oblivious adversary, it suffices to upper bound the worst-case pseudo-regret under a deterministic oblivious adversary.

2 EXP3

We will now study an algorithm called EXP3; this algorithm obtains low pseudo-regret (and hence low expected regret against an oblivious adversary). The idea of EXP3 is to try to run Hedge, but this is not actually possible since Learner only observes the loss of the arm it pulls in each round. EXP3 instead maintains *importance-weighted loss estimates* of the losses based on the information it observes, and it runs Hedge on these loss estimates instead.

Let us first look at how EXP3 forms its loss estimates. Similar to Hedge and the exponentially weighted average forecaster, in each round EXP3 maintains a distribution over actions. In round t, EXP3 pulls an arm I_t drawn from a distribution p_t. For each arm $i \in [K]$, it then uses the following importance-weighted loss estimate of $\ell_{i,t}$:

$$\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} \cdot \mathbf{1}\left[I_t = i\right].$$

The reason for this choice of loss estimate is that $\tilde{\ell}_{i,t}$ is an unbiased estimator of $\ell_{i,t}$, since

$$\mathsf{E}_{I_t \sim p_t}\left[\tilde{\ell}_{i,t}\right] = \sum_{j=1}^{K} p_{j,t} \frac{\ell_{i,t}}{p_{i,t}} \cdot \mathbf{1}\left[j = i\right] = \ell_{i,t}. \tag{6}$$

Let $\widetilde{L}_{i,t} = \sum_{s=1}^{t} \tilde{\ell}_{i,s}$ denote the cumulative importance-weighted loss of arm i. The full algorithm is shown below.

Algorithm 1: EXP3

Input: $\eta > 0$

Set $p_{j,1} = \frac{1}{K}$ for $j = 1, \ldots, K$

for $t = 1 \to T$ **do**

Draw arm I_t according to probability distribution p_t

Pull arm I_t and observe loss $\ell_{I_t,t}$

For $i \in [K]$, compute loss estimate $\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} \cdot \mathbf{1}[I_t = i]$

For $i \in [K]$, set $p_{i,t+1} = \dfrac{e^{-\eta \widetilde{L}_{i,t}}}{\sum_{j=1}^{K} e^{-\eta \widetilde{L}_{j,t}}}$

end

The next result implies an upper bound on the pseudo-regret of EXP3. The result is most useful for oblivious adversaries (although technically it also holds for non-oblivious adversaries).

Theorem 1. *For any $\eta > 0$ and any arm $i \in [K]$, EXP3's expected regret against arm i is bounded as*

$$\mathsf{E}\left[\hat{L}_T - L_{i,t}\right] \leq \frac{\log K}{\eta} + \frac{TK\eta}{2}.$$

In particular, the choice $\eta = \sqrt{\frac{2 \log K}{KT}}$ gives the bound $\sqrt{2TK \log K}$.

Note that since we the upper bound holds for all $i \in [K]$, it also holds when taking the maximum over $i \in [K]$ (which is the pseudo-regret).

We will use the following lemma to prove Theorem 1.

Lemma 1. *Let X be a nonnegative random variable. Then*

$$\log \mathsf{E}\left[e^{-X}\right] + \mathsf{E}[X] \leq \mathsf{E}\left[\frac{X^2}{2}\right]$$

Proof. We first use the inequality $\log x \leq x - 1$, which gives

$$\log \mathsf{E}\left[e^{-X}\right] + \mathsf{E}[X] \leq \mathsf{E}\left[e^{-X} - 1 + X\right]. \tag{7}$$

Next, we use the following inequality[1]:

$$e^{-x} - 1 + x \leq \frac{x^2}{2} \qquad \text{for } x \geq 0. \tag{8}$$

Applying (8), the right-hand side of (7) is at most $\mathsf{E}\left[\frac{X^2}{2}\right]$. □

[1] To see why (8) holds, observe that

$$e^{-x} - 1 + x - \frac{x^2}{2} = -\frac{x^3}{3!} + \frac{x^4}{4!} - \frac{x^5}{5!} + \ldots.$$

The right-hand side is zero for $x = 0$. It is enough to verify that the first derivative is nonpositive for all $x \geq 0$. For this, observe that the first derivative also is zero for $x = 0$, and so it is enough to verify that the second derivative is nonpositive for all $x \geq 0$. *For this*, observe that the third derivative is equal to $-e^{-x}$, which is of course nonpositive for all $x \geq 0$. Thus, going backwards, all of the required conditions are satisfied, and (8) indeed holds.

Proof (of Theorem 1). Consider an arbitrary arm $i \in [K]$. For any $t \in [T]$, let

$$\widetilde{m}_t := -\frac{1}{\eta} \log \mathsf{E}_{j \sim p_t} \left[e^{-\eta \tilde{\ell}_{j,t}} \right]$$

be the mix loss based on the loss estimates, and define $\widetilde{M}_t := \sum_{s=1}^{t} \widetilde{m}_s$. Just like our analysis of the Exponentially Weighted Average (EWA) Forecaster, we use the decomposition

$$\sum_{t=1}^{T} \langle \tilde{\ell}_t, p_t \rangle - \sum_{t=1}^{T} \tilde{\ell}_{i,t} = \left(\sum_{t=1}^{T} \langle \tilde{\ell}_t, p_t \rangle - \widetilde{M}_T \right) + \left(\widetilde{M}_T - \sum_{t=1}^{T} \tilde{\ell}_{i,t} \right). \tag{9}$$

The idea of the proof is to show that the expectation of the left-hand side is equal to the pseudo-regret against arm i and to then bound the expectation of the right-hand side. We do these steps in sequence. For any $t \in [T]$, let $\mathcal{F}_{t-1}$ be the history formed by the first $t-1$ rounds. From the law of total expectation and the unbiased property of the importance-weighted loss estimators respectively, we have

$$\begin{aligned}
\mathsf{E} \left[\sum_{t=1}^{T} \langle \tilde{\ell}_t, p_t \rangle - \sum_{t=1}^{T} \tilde{\ell}_{i,t} \right] &= \mathsf{E} \left[\sum_{t=1}^{T} \left\langle \mathsf{E} \left[\tilde{\ell}_t \mid \mathcal{F}_{t-1} \right], p_t \right\rangle - \sum_{t=1}^{T} \mathsf{E} \left[\tilde{\ell}_{i,t} \mid \mathcal{F}_{t-1} \right] \right] \\
&= \mathsf{E} \left[\sum_{t=1}^{T} \langle \ell_t, p_t \rangle - \sum_{t=1}^{T} \ell_{i,t} \right] \\
&= \mathsf{E} \left[\hat{L}_T - L_{i,T} \right]
\end{aligned}$$

We now control the right-hand side of (9), beginning with the second term. Recall that Lemma 2 of Lecture 15 showed that for the EWA Forecaster (which includes Hedge as a special case), the cumulative mix loss is not much larger than the cumulative loss of any expert. That lemma did not require losses to be in the range $[0, 1]$, so the same argument works even when we have loss estimates in place of the actual losses. Therefore,

$$\widetilde{M}_T - \sum_{t=1}^{T} \tilde{\ell}_{i,t} \leq \frac{\log K}{\eta}. \tag{10}$$

It remains to upper bound the first term on the right-hand side of (9), i.e., $\sum_{t=1}^{T} \langle \tilde{\ell}_t, p_t \rangle - \widetilde{M}_T$. We cannot use Hoeffding's Lemma as we did with the EWA Forecaster because that lemma requires control on the range of the losses, whereas here we use loss estimates whose range can be very large. So, we proceed a different way. We will bound $\langle \tilde{\ell}_t, p_t \rangle - \widetilde{m}_t$ for a fixed round t. To this end, observe that

$$\langle \tilde{\ell}_t, p_t \rangle - \widetilde{m}_t = \frac{1}{\eta} \left(\log \mathsf{E}_{j \sim p_t} \left[e^{-\eta \tilde{\ell}_{j,t}} \right] + \mathsf{E}_{j \sim p_t} \left[\eta \tilde{\ell}_{j,t} \right] \right).$$

Therefore, we can apply Lemma 1 with $X = \eta \tilde{\ell}_{j,t}$ to get

$$\langle \tilde{\ell}_t, p_t \rangle - \widetilde{m}_t \leq \frac{1}{\eta} \mathsf{E}_{j \sim p_t} \left[\frac{\left(\eta \tilde{\ell}_{j,t} \right)^2}{2} \right] = \frac{\eta}{2} \mathsf{E}_{j \sim p_t} \left[\left(\tilde{\ell}_{j,t} \right)^2 \right] \tag{11}$$

Applying (10) and (11) in (9) gives

$$\begin{aligned}
\sum_{t=1}^{T}\langle\tilde{\ell}_t, p_t\rangle - \sum_{t=1}^{T}\tilde{\ell}_{i,t} &\leq \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathsf{E}_{j\sim p_t}\left[(\tilde{\ell}_{j,t})^2\right] \\
&= \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{j=1}^{K} p_{j,t}\left(\frac{\ell_{j,t}\,\mathbf{1}\left[I_t = j\right]}{p_{j,t}}\right)^2 \\
&= \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{j=1}^{K}\frac{\ell_{j,t}^2\,\mathbf{1}\left[I_t = j\right]}{p_{j,t}} \\
&\leq \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{j=1}^{K}\frac{\mathbf{1}\left[I_t = j\right]}{p_{j,t}}.
\end{aligned}$$

Now, taking the expectation on both sides of the above inequality and noting the equality $\mathsf{E}\left[\frac{\mathbf{1}[I_t=j]}{p_{j,t}}\right] = 1$, we have

$$\begin{aligned}
\mathsf{E}\left[\hat{L}_T - L_{i,T}\right] &= \mathsf{E}\left[\sum_{t=1}^{T}\langle\tilde{\ell}_t, p_t\rangle - \sum_{t=1}^{T}\tilde{\ell}_{j,t}\right] \\
&\leq \frac{\log K}{\eta} + \mathsf{E}\left[\frac{\eta}{2}\sum_{t=1}^{T}\sum_{j=1}^{K}\frac{\mathbf{1}\left[I_t = j\right]}{p_{j,t}}\right] \\
&\leq \frac{\log K}{\eta} + \frac{TK\eta}{2}.
\end{aligned}$$

Taking the choice $\eta = \sqrt{\frac{2\log K}{TK}}$ gives the regret bound $\sqrt{\frac{TK\log K}{2}}$. □