Machine Learning Theory (CSC 482A/581B) - Lecture 22

Nishant Mehta

1 The non-stochastic multi-armed bandit problem

In the non-stochastic multi-armed bandit problem, also called the adversarial multi-armed bandit problem, the losses are no longer assumed to be generated in an i.i.d. stochastic fashion. Instead, the goal is to obtain low regret even when the sequence of losses is generated in an arbitrary way, whether this be from an oblivious opponent (who possibly randomizes) or a non-oblivious (also called "reactive") adversary that selects $\ell_t = (\ell_{1,t}, \dots, \ell_{K,t})$ with knowledge of the previous plays of the learning algorithm.

Let I_t denote the arm played by the learning algorithm in round t. With this notation, the regret of the learning algorithm is

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

The capitalization of the variable I_t is intentional: any learning algorithm that obtains low regret must necessarily randomize, even if the adversary is only oblivious, and so 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 $\ell_1, \dots, \ell_{t-1}$, the learning algorithm always plays a fixed action I_t. Then in round t, the adversary 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

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

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

Moreover, the adversary 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 the adversary 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, but this is beyond the scope of this course.

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 the adversary 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's see why the second observation is true. First, suppose that the adversary is deterministic; 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 adversaries, the pseudo-regret is equal to the expected regret. Next, suppose that the adversary is oblivious but might also randomize. Let $\mathbf{B}$ denote the randomization of the adversary. Then, since the adversary 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 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

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

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 adversary.

2 Exp3

We will now study an algorithm called Exp3; thsi algorithm obtains low pseudo-regret (and hence low expected regret against an oblivious adversary). The idea of Exp3 is to try to use an exponential weights-type algorithm, but the usual exponential weight updates are not directly possible since we only observe the loss of the arm we pull in each round. Exp3 instead maintains *estimates* of the losses based on the information it observes, and it updates its weights using these loss estimates instead.

Let's 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 estimates the loss $\ell_{i,t}$ as

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

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

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

The full algorithm is shown below. We use the notation $\tilde{L}_{i,t} = \sum_{s=1}^{t} \tilde{\ell}_{i,s}$.

Algorithm 1. Exp3

Given: $\eta > 0$

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

For $t = 1, \dots, T$:

1. Draw arm I_t according to probability distribution p_t
2. For $i \in [K]$, compute loss estimate $\tilde{\ell}_{i,t} = \frac{\ell_{i,t}}{p_{i,t}} \cdot \mathbf{1}_{\{I_t = i\}}$
3. For $i \in [K]$, set $p_{i,t} = \dfrac{e^{-\eta \tilde{L}_{i,t}}}{\sum_{j=1}^{K} e^{-\eta \tilde{L}_{i,t}}}$.

The next result upper bounds the pseudo-regret of Exp3.

Theorem 1. *If Exp3 is run with the learning rate* $\eta = \sqrt{\frac{2 \log K}{TK}}$*, then for any adversary,*

$$\bar{R}_T \leq \sqrt{2TK \log K}.$$

From (5), Exp3 enjoys the same upper bound for the expected regret $\mathsf{E}[R_T]$ under any oblivious adversary.

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 $\mathsf{E}\left[\frac{X^2}{2}\right]$. □

Proof of Theorem 1. To upper bound the pseudo-regret, it suffices to upper bound the expected regret against an arbitrary arm $k \in [K]$, i.e. to upper bound:

$$\mathsf{E}\left[\sum_{t=1}^{T} \ell_{I_t,t} - \sum_{t=1}^{T} \ell_{k,t}\right]$$

We will proceed by deriving an upper bound on $\sum_{t=1}^{T} \ell_{I_t,t}$, and near the end of the proof we will take the full expectation of this quantity.

The first step is to express each instantaneous loss in terms of a certain expectation of a loss estimate. To this end, observe that for each $t \in [T]$,

$$\mathsf{E}_{i \sim p_t}\left[\tilde{\ell}_{i,t}\right] = \ell_{I_t,t},$$

and so the cumulative loss of Exp3 can be expressed as

$$\sum_{t=1}^{T} \mathsf{E}_{i \sim p_t}\left[\tilde{\ell}_{i,t}\right].$$

The next step is to upper bound each of the individual terms in this summation. We begin with the trivial, yet useful, observation that[2]

$$\eta\, \mathsf{E}_{i \sim p_t}[\tilde{\ell}_{i,t}] = \log \mathsf{E}_{i \sim p_t}\left[e^{-\eta \tilde{\ell}_{i,t}}\right] + \eta\, \mathsf{E}_{i \sim p_t}[\tilde{\ell}_{i,t}] - \log \mathsf{E}_{i \sim p_t}\left[e^{-\eta \tilde{\ell}_{i,t}}\right].$$

Applying Lemma 1 for $X = \eta \tilde{\ell}_{i,t}$ and dividing by η, we have

$$\begin{aligned} \mathsf{E}_{i \sim p_t}[\tilde{\ell}_{i,t}] &\leq \frac{\eta}{2}\, \mathsf{E}_{i \sim p_t}\left[(\tilde{\ell}_{i,t})^2\right] - \frac{1}{\eta} \log \mathsf{E}_{i \sim p_t}\left[e^{-\eta \tilde{\ell}_{i,t}}\right] \\ &\leq \frac{\eta}{2 p_{I_t,t}} - \frac{1}{\eta} \log \mathsf{E}_{i \sim p_t}\left[e^{-\eta \tilde{\ell}_{i,t}}\right]. \end{aligned} \tag{9}$$

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

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

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.

[2] Note that $\eta\, \mathsf{E}[X] + \log \mathsf{E}\left[e^{-\eta X}\right] = \log \mathsf{E}\left[e^{-\eta(X - \mathsf{E}[X])}\right]$, the cumulant generating function of the centered random variable $X - \mathsf{E}[X]$, evaluated at η. This offers another interpretation of Lemma 1.

We now focus on the second term in (9). From the definition of p_t, we have

$$\log \mathsf{E}_{i\sim p_t}\left[e^{-\eta\tilde{\ell}_{i,t}}\right] = \log \frac{\sum_{i=1}^K e^{-\eta\tilde{L}_{i,t-1}} e^{-\eta\tilde{\ell}_{i,t}}}{\sum_{i=1}^K e^{-\eta\tilde{L}_{i,t-1}}} = \log \frac{\sum_{i=1}^K e^{-\eta\tilde{L}_{i,t}}}{\sum_{i=1}^K e^{-\eta\tilde{L}_{i,t-1}}}.$$

Therefore, summing (9) over $t \in [T]$, the second term of (9) becomes a telescoping series, yielding

$$\begin{aligned}
\sum_{t=1}^T \mathsf{E}_{i\sim p_t}[\tilde{\ell}_{i,t}] &\le \sum_{t=1}^T \frac{\eta}{2p_{I_t,t}} + \frac{1}{\eta}\log\sum_{i=1}^K e^{-\eta\tilde{L}_{i,0}} - \frac{1}{\eta}\log\sum_{i=1}^K e^{-\eta\tilde{L}_{i,T}} \\
&= \sum_{t=1}^T \frac{\eta}{2p_{I_t,t}} + \frac{\log K}{\eta} - \frac{1}{\eta}\log\sum_{i=1}^K e^{-\eta\tilde{L}_{i,T}} \\
&\le \sum_{t=1}^T \frac{\eta}{2p_{I_t,t}} + \frac{\log K}{\eta} - \frac{1}{\eta}\log e^{-\eta\tilde{L}_{k,T}} \\
&= \sum_{t=1}^T \frac{\eta}{2p_{I_t,t}} + \frac{\log K}{\eta} + \sum_{t=1}^T \tilde{\ell}_{k,t}.
\end{aligned}$$

Finally, taking the expectation yields

$$\begin{aligned}
\mathsf{E}\left[\sum_{t=1}^T \ell_{I_t,t}\right] &= \mathsf{E}\left[\sum_{t=1}^T \mathsf{E}_{i\sim p_t}\left[\tilde{\ell}_{i,t}\right]\right] \\
&\le \mathsf{E}\left[\sum_{t=1}^T \frac{\eta}{2p_{I_t,t}} + \frac{\log K}{\eta} + \sum_{t=1}^T \tilde{\ell}_{k,t}\right] \\
&= \mathsf{E}\left[\sum_{t=1}^T \mathsf{E}_{I_t\sim p_t}\left[\frac{\eta}{2p_{I_t,t}}\right] + \frac{\log K}{\eta} + \sum_{t=1}^T \mathsf{E}_{I_t\sim p_t}\left[\tilde{\ell}_{k,t}\right]\right] \\
&= \mathsf{E}\left[\frac{TK\eta}{2} + \frac{\log K}{\eta} + \sum_{t=1}^T \ell_{k,t}\right],
\end{aligned}$$

where the first equality is from the law of total expectation and the second equality follows from direct computation of $\mathsf{E}_{I_t\sim p_t}\left[\frac{1}{p_{I_t,t}}\right] = K$ and the fact that $\tilde{\ell}_{k,t}$ is an unbiased estimator of $\ell_{k,t}$ (recall (6)). The result follows by plugging in the value $\eta = \sqrt{\frac{2\log K}{TK}}$. □

Theorem 1 is not directly useful when the time horizon T is unknown. However, by using the doubling trick, one can obtain the same upper bound with a larger constant. Alternatively, Exp3 can be run using a time-varying learning rate of $\eta_t = \sqrt{\frac{\log K}{tk}}$ to yield pseudo-regret at most $2\sqrt{TK\log K}$. The proof becomes somewhat more involved; if you are interested, see Theorem 3.1 of Bubeck and Cesa-Bianchi (2012).

References

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.