Lecture 12

Instructor: Haipeng Luo

1 The Multi-armed Bandit Problem

All the topics we have discussed so far consider problem with full information feedback. Staring from this lecture, we will move on to the more challenging settings with partial information feedback. The classic example of such problems is the *multi-armed bandit* problem [Lai and Robbins, 1985], and here we discuss an adversarial version introduced in [Auer et al., 2002].

The problem models the situation where a gambler sequentially pull the arm of one of the slot machines in a casino, with the hope of maximizing reward. A slot machine is sometimes called a "one-armed bandit", and hence the name multi-armed bandit for this problem. Formally, there are K arms/actions available for a learner, and at each time $t = 1, \ldots, T$,

1. the learner picks an action $a_t \in [K]$ while simultaneously the environment decides the loss vector $\ell_t \in [0,1]^K$,
2. the learner then suffers and observes (only) the loss $\ell_t(a_t)$.

Clearly, this is simply a partial information version of the expert problem, with the difference being that the learner has to actually pick one action at each round and then observe only the loss for this action but not the whole loss vector ℓ_t. For convention, we move from the notation i and N to a and K to denote a specific action and the total number of actions respectively.

For simplicity we only consider oblivious environment and thus one can equivalently think of the loss vectors as generated ahead of time before the game starts (possibly randomly though). We measure the algorithm's performance by the expected regret

$$\mathbb{E}\left[\mathcal{R}_T\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \min_{a\in[K]} \sum_{t=1}^{T} \ell_t(a),$$

where the expectation is with respect to the randomness of the algorithm.

The challenge of this problem (or in general all partial information problem) is the well-known exploitation-exploration tradeoff. Indeed, on one hand, it's tempting to pick actions that have suffered small losses before (exploitation), but on the other hand, there is also an incentive to pick other actions just to see whether they can admit even smaller losses (exploration).

But since the problem is so close to the expert problem, let's first see whether we can somehow use an expert algorithm to solve it. The obvious obstacle is that we do not have the whole loss vector to feed to an expert algorithm. However, suppose we pick a_t according to a distribution p_t, then we can construct an estimator for the loss vector in the following way

$$\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)}\mathbf{1}\{a = a_t\} = \begin{cases} \frac{\ell_t(a_t)}{p_t(a_t)} & \text{if } a = a_t, \\ 0 & \text{else.} \end{cases}$$

This simple trick is called inverse propensity score weighting or simply importance weighting. Apparently the estimator is computable using the available information, and more importantly, it is unbiased: for any $a \in [K]$,

$$\mathbb{E}_t[\widehat{\ell}(a)] = (1 - p_t(a)) \times 0 + p_t(a)\frac{\ell_t(a)}{p_t(a)} = \ell_t(a)$$

where $\mathbb{E}_t[\cdot]$ is the conditional expectation with respect to the random draw of a_t given the past. Therefore, since we only care about expected regret (at least for now), it seems like we can simply use the prediction of an arbitrary expert algorithm p_t to draw a_t, and then feed $\widehat{\ell}_t$ to the algorithm. Indeed, we have for any $a \in [K]$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t(a_t)\right] - \sum_{t=1}^{T} \ell_t(a) = \mathbb{E}\left[\sum_{t=1}^{T} \left\langle p_t, \widehat{\ell}_t \right\rangle - \sum_{t=1}^{T} \widehat{\ell}_t(a)\right]$$

where the last term is exactly the (expected) regret of the expert algorithm. We have showed that the optimal regret for the expert problem is $\mathcal{O}(\sqrt{T \ln K})$. Does this mean we have come up with a simple algorithm for the multi-armed bandit with regret $\mathcal{O}(\sqrt{T \ln K})$?

The answer is no – what we missed in the above argument is the fact that the range of the losses that the expert algorithm receives is no longer in $[0, 1]$! In fact, it could potentially be very large due to the importance weighting and thus the regret is no longer just $\mathcal{O}(\sqrt{T \ln K})$. As a simple fix, we can try to enforce a lower bound on the importance weight by doing a small amount of uniform exploration

$$p_t = (1 - \alpha)\widehat{p}_t + \frac{\alpha}{K}\mathbf{1} \tag{1}$$

where $\widehat{p}_t$ is now the prediction of the expert algorithm, $\mathbf{1}$ is the all-one vector, and α is some parameter to be specified later. Then clearly we have $\widehat{\ell}(a) \leq K/\alpha$ and thus if we feed the expert algorithm with $\frac{\alpha}{K}\widehat{\ell}_t \in [0, 1]^K$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \left\langle p_t, \widehat{\ell}_t \right\rangle - \sum_{t=1}^{T} \widehat{\ell}_t(a)\right] &\leq \frac{\alpha}{K}\mathbb{E}\left[\sum_{t=1}^{T} \widehat{\ell}_t(a_t)\right] + \mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{p}_t, \widehat{\ell}_t \right\rangle - \sum_{t=1}^{T} \widehat{\ell}_t(a)\right] \\
&\leq \alpha T + \frac{K}{\alpha}\mathbb{E}\left[\sum_{t=1}^{T} \left\langle \widehat{p}_t, \frac{\alpha}{K}\widehat{\ell}_t \right\rangle - \sum_{t=1}^{T} \frac{\alpha}{K}\widehat{\ell}_t(a)\right] \\
&= \mathcal{O}\left(\alpha T + \frac{K}{\alpha}\sqrt{T \ln K}\right)
\end{aligned}$$

where the second step is by $\mathbb{E}_t[\widehat{\ell}_t(a_t)] = \sum_{a=1}^{K} p_t(a)\frac{\ell_t(a)}{p_t(a)} = \sum_{a=1}^{K} \ell_t(a) \leq K$ and the last step is by applying the regret bound of the expert algorithm. Finally, by picking the optimal α we achieve a regret bound of order $\mathcal{O}(T^{\frac{3}{4}} K^{\frac{1}{2}} (\ln K)^{\frac{1}{4}})$, which is much larger than the optimal bound for the full information setting. Therefore, although the importance weighted estimator is unbiased and we do only care about regret in expectation, the range or really the variance of the estimator still matters.

2 The Exp3 Algorithm

Can we do better than the approach discussed above? It turns out that the answer is yes, and the solution is simply by using Hedge as the expert algorithm in the above reduction, *without even mixing the uniform distribution*. To see this, note that the potential-based proof of Hedge does not use the fact that the losses are in $[0, 1]$ at all to arrive at the following

$$\sum_{t=1}^{T} \left\langle p_t, \widehat{\ell}_t \right\rangle - \sum_{t=1}^{T} \widehat{\ell}_t(a) \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a)\widehat{\ell}_t(a)^2. \tag{2}$$

Noting that the variance (or rather the second moment) of the estimator is $\mathbb{E}_t[\widehat{\ell}_t(a)^2] = \frac{\ell_t(a)^2}{p_t(a)}$, we continue with

$$\mathbb{E}\left[\sum_{t=1}^{T} \left\langle p_t, \widehat{\ell}_t \right\rangle - \sum_{t=1}^{T} \widehat{\ell}_t(a)\right] \leq \frac{\ln K}{\eta} + \eta\left[\sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a)\frac{\ell_t(a)^2}{p_t(a)}\right] \leq \frac{\ln K}{\eta} + TK\eta,$$

which means with the optimal tuning $\eta = \sqrt{(\ln K)/(TK)}$ the regret is only $\mathcal{O}(\sqrt{TK \ln K})$, much better than the previous $\mathcal{O}(T^{\frac{3}{4}})$ bound! This is yet another example of the power of adaptive regret

Algorithm 1: Exp3

Input: learning rate $\eta > 0$

Initialization: let $\widehat{L}_0$ be the all-zero vector

for $t = 1, \dots, T$ **do**

compute $p_t \in \Delta(K)$ such that $p_t(a) \propto \exp(-\eta \widehat{L}_{t-1}(a))$

play $a_t \sim p_t$ and observe its loss $\ell_t(a_t)$

update $\widehat{L}_t = \widehat{L}_{t-1} + \widehat{\ell}_t$ where $\widehat{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)} \mathbf{1}\{a = a_t\},\ \forall a \in [K]$

bounds, and it is in fact pretty magical that bound (2) can automatically deal with the large variance issue of the estimators.

This algorithm (summarized in Algorithm 1) is called Exp3 (which stands for Exponential-weight for Exploration and Exploitation) is the first and arguably most important algorithm for adversarial multi-armed bandit. Its regret bound is summarized in the following theorem for completeness.

Theorem 1. *With the optimal tuning Exp3 ensures* $\mathbb{E}[\mathcal{R}_T] = \mathcal{O}(\sqrt{TK \ln K})$.

It is worth noting that although there is no explicit exploration (like Eq. (1)) in Exp3, the algorithm is in fact doing some implicit exploration. Indeed, whenever an arm a_t is pulled (maybe due to exploitation), its weight for the next round is always not increased no matter what the loss vector ℓ_t is, which will then encourage the algorithm to explore other actions next round. This is due to the structure of the estimator $\widehat{\ell}_t$ so that only the picked action a_t could have non-zero loss, while all the other actions have estimated loss 0.

3 Lower Bounds

The regret bound of Exp3 is showing that the price of bandit feedback is only a $\sqrt{K}$ factor compared to full information feedback (ignoring logarithmic terms). Is this optimal?

Intuitively it should be. Consider the following very informal argument. Suppose the losses are all generated independently and uniformly from $\{0, 1\}$. For any fixed algorithm, there must be an arm that is pulled no more than T/K times by this algorithm. Now suppose the environment is modified so that the loss of this arm follows a Bernoulli distribution with parameter $1/2 - \sqrt{K/T}$, which is not distinguishable from the uniform distribution information-theoretically with only T/K samples. Then the algorithm should not be aware of this change and still pull this arm no more than T/K rounds, leading to an expected regret $(T - T/K)\sqrt{K/T} \approx \sqrt{TK}$.

The following theorem makes the argument above formal with a probabilistic argument, similar to the lower bound proof for the expert problem.

Theorem 2. *For any multi-armed bandit algorithm $\mathcal{A}$, there exists a sequence of loss vectors s.t.*

$$\mathbb{E}_{\mathcal{A}}[\mathcal{R}_T] = \Omega(\sqrt{TK})$$

where we use $\mathbb{E}_{\mathcal{A}}[\cdot]$ to denote the expectation with respect to the randomness of $\mathcal{A}$.

Proof. Consider randomly generating an environment in the following way: first draw an action uniformly at random to be the "good" action; then for all $t \in [T]$ and $a \in [K]$ independently generate $\ell_t(a)$ whose distribution is a Bernoulli with parameter $1/2 - \epsilon$ if a is the good action (ϵ to be specified later), or uniform on $\{0, 1\}$ otherwise. Let $\mathbb{E}_\star[\cdot]$ be the expectation with respect to the random draw of such environment. Our goal is to prove

$$\mathbb{E}_\star[\mathbb{E}_{\mathcal{A}}[\mathcal{R}_T]] = \Omega(\sqrt{TK}) \tag{3}$$

which clearly implies the theorem.

The first step to prove Eq. (3) is to realize that $\mathbb{E}_\star[\mathbb{E}_{\mathcal{A}}[\mathcal{R}_T]] = \mathbb{E}_{\mathcal{A}}[\mathbb{E}_\star[\mathcal{R}_T]]$ and thus it is enough to show that for any deterministic algorithm, $\mathbb{E}_\star[\mathcal{R}_T]$ has the same lower bound. Note that for a deterministic algorithm, a_t is completely determined by $\widetilde{\ell}_{1:t-1}$, a shorthand for $\ell_1(a_1), \dots, \ell_{t-1}(a_{t-1})$.

Now let $\mathbb{E}_a[\cdot]$ denote the conditional expectation given that the good action is a, we have

$$\begin{aligned}
\mathbb{E}_\star[\mathcal{R}_T] &= \frac{1}{K}\sum_{a=1}^{K}\mathbb{E}_a\left[\sum_{t=1}^{T}\ell_t(a_t) - \min_{a^\star\in[K]}\sum_{t=1}^{T}\ell_t(a^\star)\right] \\
&\geq \frac{1}{K}\sum_{a=1}^{K}\mathbb{E}_a\left[\sum_{t=1}^{T}\ell_t(a_t) - \sum_{t=1}^{T}\ell_t(a)\right] = \frac{1}{K}\sum_{a=1}^{K}\mathbb{E}_a\left[\sum_{t:a_t\neq a}(\ell_t(a_t)-\ell_t(a))\right] \\
&\geq \frac{1}{K}\sum_{a=1}^{K}\mathbb{E}_a\left[\sum_{t:a_t\neq a}\epsilon\right] = \epsilon\left(T - \frac{1}{K}\sum_{a=1}^{K}\mathbb{E}_a\left[n_a\right]\right) \qquad (4)
\end{aligned}$$

where n_a is the number of times that a is picked by the algorithm. To upper bound the term $\mathbb{E}_a[n_a]$, we imagine a reference environment where every loss is an independent and uniform draw from $\{0,1\}$, and let $\mathbb{E}_0[\cdot]$ denote the corresponding expectation. Noting that n_a is a function of $\widetilde{\ell}_{1:T}$, with $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_K$ being the probability distribution of $\widetilde{\ell}_{1:T}$ under the corresponding environment, we can relate $\mathbb{E}_a[n_a]$ and $\mathbb{E}_0[n_a]$ as

$$\mathbb{E}_a[n_a] - \mathbb{E}_0[n_a] = \sum_{\widetilde{\ell}_{1:T}} n_a\left(\mathbb{P}_a(\widetilde{\ell}_{1:T}) - \mathbb{P}_0(\widetilde{\ell}_{1:T})\right) \leq T\sum_{\widetilde{\ell}_{1:T}}\left|\mathbb{P}_a(\widetilde{\ell}_{1:T}) - \mathbb{P}_0(\widetilde{\ell}_{1:T})\right| = T\left\|\mathbb{P}_a - \mathbb{P}_0\right\|_1$$

which by Pinskers inequality is bounded by $\sqrt{2\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_a)}$. We compute the KL term as follows:

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_a) &= \sum_{\widetilde{\ell}_{1:T}}\mathbb{P}_0(\widetilde{\ell}_{1:T})\ln\left(\frac{\mathbb{P}_0(\widetilde{\ell}_{1:T})}{\mathbb{P}_a(\widetilde{\ell}_{1:T})}\right) = \sum_{\widetilde{\ell}_{1:T}}\mathbb{P}_0(\widetilde{\ell}_{1:T})\ln\left(\frac{\prod_{t=1}^{T}\mathbb{P}_0(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\prod_{t=1}^{T}\mathbb{P}_a(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{t=1}^{T}\sum_{\widetilde{\ell}_{1:t}}\mathbb{P}_0(\widetilde{\ell}_{1:t})\ln\left(\frac{\mathbb{P}_0(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}_a(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) = \sum_{t=1}^{T}\sum_{\widetilde{\ell}_{1:t}:a_t=a}\mathbb{P}_0(\widetilde{\ell}_{1:t})\ln\left(\frac{\mathbb{P}_0(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}_a(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \sum_{t=1}^{T}\sum_{\widetilde{\ell}_{1:t-1}:a_t=a}\mathbb{P}_0(\widetilde{\ell}_{1:t-1})\sum_{\widetilde{\ell}_t\in\{0,1\}}\mathbb{P}_0(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})\ln\left(\frac{\mathbb{P}_0(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}{\mathbb{P}_a(\widetilde{\ell}_t|\widetilde{\ell}_{1:t-1})}\right) \\
&= \frac{1}{2}\sum_{t=1}^{T}\mathbb{P}_0(a_t=a)\left(\ln\frac{1/2}{1/2+\epsilon} + \ln\frac{1/2}{1/2-\epsilon}\right) = \frac{\mathbb{E}_0[n_a]}{2}\ln\left(\frac{1}{1-4\epsilon^2}\right).
\end{aligned}$$

Therefore, we have by $\sum_{a=1}^{K}\mathbb{E}_0[n_a] = T$ and Cauchy-Schwarz inequality

$$\sum_{a=1}^{K}\mathbb{E}_a[n_a] \leq \sum_{a=1}^{K}\mathbb{E}_0[n_a] + T\sum_{a=1}^{K}\sqrt{\mathbb{E}_0[n_a]\ln\left(\frac{1}{1-4\epsilon^2}\right)} \leq T + T\sqrt{KT\ln\left(\frac{1}{1-4\epsilon^2}\right)}.$$

Plugging the above back to Eq. (4) shows

$$\mathbb{E}_\star[\mathcal{R}_T] \geq \epsilon T\left(1 - \frac{1}{K} - \sqrt{\frac{T}{K}\ln\left(\frac{1}{1-4\epsilon^2}\right)}\right) = \Omega\left(\epsilon T\left(1 - \epsilon\sqrt{\frac{T}{K}}\right)\right),$$

which proves Eq. (3) with the optimal ϵ and finishes the proof. □

Therefore we see that Exp3 is almost worst-case optimal. In the next lecture we will discuss algorithms that are exactly optimal up to constants.

References

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.