Algorithms and Uncertainty, Summer 2021	Lecture 20 (4 pages)
No-Regret Learning: Multi-Armed Bandits 1	
Thomas Kesselheim	Last Update: July 7, 2021

1 Last Lecture

Let us first summarize what we have seen in the last lecture. We consider an online learning setting, in which our algorithm has n choices in each step, each choice corresponds to an *expert*.

First an adversary chooses a sequence of cost vectors $\ell^{(1)}, \ldots, \ell^{(T)}$. Then, in step t, the algorithm first chooses one of the n experts (possibly in a randomized way), which we call I_t. Then the algorithm gets to know the entire vector $\ell^{(t)}$.

If $\ell_i^{(t)} \in [0,1]$, we showed that Multiplicative Weights (MW) is a randomized algorithm (with parameter η) that guarantees

$$\mathbf{E}\left[\sum_{t=1}^{T} \ell_{I_t}^{(t)}\right] \leq (1+\eta) \min_i \sum_{t=1}^{T} \ell_i^{(t)} + \frac{\ln n}{\eta} \quad .$$

By setting $\eta = \sqrt{\frac{\ln n}{T}}$, we get

$$\mathbf{E}\left[\sum_{t=1}^{T} \ell_{I_t}^{(t)}\right] \leq \min_i \sum_{t=1}^{T} \ell_i^{(t)} + \sqrt{\frac{\ln n}{T}} \min_i \sum_{t=1}^{T} \ell_i^{(t)} + \sqrt{T \ln n} \leq \min_i \sum_{t=1}^{T} \ell_i^{(t)} + 2\sqrt{T \ln n} \quad .$$

The quantity $\mathrm{Regret}^{(T)} = \mathbf{E}\left[\sum_{t=1}^{T} \ell_{I_t}^{(t)}\right] - \min_i \sum_{t=1}^{T} \ell_i^{(t)}$ is called the *(external) regret* on the sequence. Multiplicative Weights guarantees that the regret is always bounded by $2\sqrt{T \ln n}$.

An algorithm that guarantees $\mathrm{Regret}^{(T)} = o(T)$ is called *no regret* because asymptotically the algorithm does as well as the best expert.

2 Today: Partial Feedback (Adversarial Multi-Armed Bandits)

Today, we consider again the setting that we can choose between n actions in every step. An adversary determines the sequence of cost vectors $\ell^{(1)}, \ldots, \ell^{(T)}$ in advance and it is unknown to the algorithm. We assume that $\ell_i^{(t)} \in [0,1]$ for all i and t.

In step t, the algorithm chooses one of the n actions at random by defining probabilities $p_1^{(t)}, \ldots p_n^{(t)}$. The algorithm's choice in step t is denoted by I_t. The algorithm gets to know $\ell_{I_t}^{(t)}$. The other entries of the cost vector remain unknown.

In practice often the cost or reward of alternative actions are not revealed. For example, if we run a news website, we might want to choose article headlines so as to maximize the number of clicks or shares. For each user that arrives, we can only try out one particular choice and we do not get to know how others would have performed.

Again, we are interested in a no-regret algorithm, so the algorithm should ensure that for all sequences $\ell^{(1)}, \ldots, \ell^{(T)}$, the regret

$$\mathrm{Regret}^{(T)} = \mathbf{E}\left[\sum_{t=1}^{T} \ell_{I_t}^{(t)}\right] - \min_i \sum_{t=1}^{T} \ell_i^{(t)}$$

grows sublinearly, that is, $\mathrm{Regret}^{(T)} = o(T)$.

3 Idea for a Black-Box Transformation

We will now get to know a black-box transformation to solve the bandits setting with an algorithm for the experts setting. The idea is as follows: We run an experts algorithm like Multiplicative Weights and we only give it the feedback that we have in an ingenious way. Suppose we are in round t and the algorithm chooses to play expert i with probability $p_i^{(t)}$. We do the same and get to know $\ell_{I_t}^{(t)}$. The values $\ell_i^{(t)}$ for $i \neq I_t$ are unknown to us.

The question is what feedback to return to the expert algorithm. Ideally we would want to set $\tilde{\ell}_{I_t}^{(t)} = \ell_{I_t}^{(t)} / p_{I_t}^{(t)}$ and $\tilde{\ell}_i^{(t)} = 0$ for $i \neq I_t$ and tell the experts algorithm that the feedback was $\tilde{\ell}^{(t)}$. This makes sense because $\mathbf{E}\left[\tilde{\ell}_i^{(t)}\right] = p_i^{(t)} \cdot \ell_i^{(t)} / p_i^{(t)} = \ell_i^{(t)}$, so *in expectation* the feedback is just right.

There is one thing, we have to be careful about: We will not have $\tilde{\ell}_i^{(t)} \in [0,1]$ this way, which was our assumption.

4 First Step: Cost Scaling

As a tiny first step, we observe that it is easy to extend any online learning algorithm from the setting in which $\ell_i^{(t)} \in [0,1]$ to the setting in which $\ell_i^{(t)} \in [0, \rho]$, namely by feeding $\hat{\ell}_i^{(t)} = \frac{1}{\rho} \ell_i^{(t)}$ into the algorithm.

Now, if the algorithm guarantees

$$\mathbf{E}\left[\sum_{t=1}^{T} \hat{\ell}_{I_t}^{(t)}\right] \leq \alpha \min_i \sum_{t=1}^{T} \hat{\ell}_i^{(t)} + \beta \enspace ,$$

then it also guarantees

$$\mathbf{E}\left[\sum_{t=1}^{T} \ell_{I_t}^{(t)}\right] \leq \alpha \min_i \sum_{t=1}^{T} \ell_i^{(t)} + \rho\beta \enspace .$$

In particular, the multiplicative-weights algorithm guarantees a regret of $2\sqrt{T \ln n}$ if $\ell_i^{(t)} \in [0,1]$. So, the regret will be at most $2\rho\sqrt{T \ln n}$ if $\ell_i^{(t)} \in [0, \rho]$

5 Second Step: Adding Exploration

Coming back to the idea that $\tilde{\ell}_{I_t}^{(t)} = \ell_{I_t}^{(t)} / p_{I_t}^{(t)}$ and $\tilde{\ell}_i^{(t)} = 0$ for $i \neq I_t$. Unfortunately, $p_i^{(t)}$ can be arbitrarily small, so $\tilde{\ell}_i^{(t)}$ is unbounded. We have an algorithm, which works on cost vectors between 0 and ρ. Therefore, we will increase $p_i^{(t)}$ by a small additive term to keep the numbers bounded. Our overall algorithm now looks as follows.

In step t:

- Get probability vector $p^{(t)}$ from experts algorithm.
- Set $q_i^{(t)} = (1-\gamma) p_i^{(t)} + \frac{\gamma}{n}$.
- Choose I_t based on $q^{(t)}$.
- Return $\tilde{\ell}_{I_t}^{(t)} = \ell_{I_t}^{(t)} / q_{I_t}^{(t)}$ and $\tilde{\ell}_i^{(t)} = 0$ for $i \neq I_t$ to the experts algorithm with $\rho = \frac{n}{\gamma}$.

Note that, by our assumption $\ell_i^{(t)} \in [0,1]$, it is guaranteed that $\tilde{\ell}_i^{(t)} \in [0,\rho]$ for $\rho = \frac{n}{\gamma}$.

It is also intuitive that we do not want the $p_i^{(t)}$ to become too small. If, for example, $p_i^{(t)}$ is almost 0, then we would never choose this action again and we would not recognize when it becomes much better than the other actions. So, one can also understand adding $\frac{\gamma}{n}$ as introducing some exploration to our algorithm.

6 Analysis of the Algorithm

The black-box reduction works for any experts algorithm. To keep the analysis concrete, we only consider the case in which we use Multiplicative Weights.

Theorem 20.1. *When using Multiplicative Weights as the experts algorithm, the bandits algorithm guarantees that for any sequence $\ell^{(1)}, \ldots, \ell^{(T)}$*

$$\mathbf{E}\left[\sum_{t=1}^{T} \ell_{I_t}^{(t)}\right] \leq (1+\eta) \min_i \sum_{t=1}^{T} \ell_i^{(t)} + \frac{n \ln n}{\gamma \eta} + \gamma T \ .$$

Proof. Let us first fix a choice of $I_1, \ldots, I_T$. This fixes the sequence $\tilde{\ell}^{(1)}, \ldots, \tilde{\ell}^{(T)}$ that is given to Multiplicative Weights. What would Multiplicative Weights do on this sequence? It computes probability vectors $p^{(1)}, \ldots, p^{(T)}$. These vectors have the property that

$$\sum_{t=1}^{T} \sum_{i=1}^{n} p_i^{(t)} \tilde{\ell}_i^{(t)} \leq (1+\eta) \min_i \sum_{t=1}^{T} \tilde{\ell}_i^{(t)} + \rho \frac{\ln n}{\eta} = (1+\eta) \min_i \sum_{t=1}^{T} \tilde{\ell}_i^{(t)} + \frac{n \ln n}{\gamma \eta} \ .$$

As we set $q_i^{(t)} = (1-\gamma) p_i^{(t)} + \frac{\gamma}{n}$, we also have

$$\sum_{t=1}^{T} \sum_{i=1}^{n} q_i^{(t)} \tilde{\ell}_i^{(t)} = (1-\gamma) \sum_{t=1}^{T} \sum_{i=1}^{n} p_i^{(t)} \tilde{\ell}_i^{(t)} + \frac{\gamma}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \tilde{\ell}_i^{(t)} \leq (1+\eta) \min_i \sum_{t=1}^{T} \tilde{\ell}_i^{(t)} + \frac{n \ln n}{\gamma \eta} + \frac{\gamma}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \tilde{\ell}_i^{(t)} \ .$$

So far, we kept $I_1, \ldots, I_T$ fixed. It is important to remark at this point that only our algorithm produces this "fake" sequence during the run and we tried out what Multiplicative Weights would do on the sequence. In the next step, we take the expectation over $I_1, \ldots, I_T$ on both sides.

$$\mathbf{E}\left[\sum_{t=1}^{T} \sum_{i=1}^{n} q_i^{(t)} \tilde{\ell}_i^{(t)}\right] \leq \mathbf{E}\left[(1+\eta) \min_i \sum_{t=1}^{T} \tilde{\ell}_i^{(t)} + \frac{n \ln n}{\gamma \eta} + \frac{\gamma}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \tilde{\ell}_i^{(t)}\right] \ .$$

Note that $\mathbf{E}\left[\min_i \sum_{t=1}^{T} \tilde{\ell}_i^{(t)}\right] \leq \min_i \sum_{t=1}^{T} \mathbf{E}\left[\tilde{\ell}_i^{(t)}\right]$. So, by linearity of expectation

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{E}\left[q_i^{(t)} \tilde{\ell}_i^{(t)}\right] \leq (1+\eta) \min_i \sum_{t=1}^{T} \mathbf{E}\left[\tilde{\ell}_i^{(t)}\right] + \frac{n \ln n}{\gamma \eta} + \frac{\gamma}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{E}\left[\tilde{\ell}_i^{(t)}\right] \ .$$

This inequality still talks about the fake sequence $\tilde{\ell}^{(1)}, \ldots, \tilde{\ell}^{(T)}$, but we actually want to talk about the real sequence $\ell^{(1)}, \ldots, \ell^{(T)}$.

For the term $\mathbf{E}\left[\tilde{\ell}_i^{(t)}\right]$ on the right-hand side, this is pretty easy. Let us fix $I_1 = i_1, \ldots, I_{t-1} = i_{t-1}$ arbitrarily. This fixes $q^{(t)}$ and $\mathbf{Pr}\left[I_t = i \mid I_1 = i_1, \ldots, I_{t-1} = i_{i-1}\right] = q_i^{(t)}$. So

$$\mathbf{E}\left[\tilde{\ell}_i^{(t)} \,\middle|\, I_1 = i_1, \ldots, I_{t-1} = i_{t-1}\right] = q_i^{(t)} \cdot \ell_i^{(t)} / q_i^{(t)} = \ell_i^{(t)}$$

for *any* choices of $i_1, \ldots, i_{t-1}$. So, also

$$\begin{aligned}\mathbf{E}\left[\tilde{\ell}_i^{(t)}\right] &= \sum_{i_1} \cdots \sum_{i_{t-1}} \mathbf{Pr}\left[I_1 = i_1, \ldots, I_{t-1} = i_{t-1}\right] \mathbf{E}\left[\tilde{\ell}_i^{(t)} \;\middle|\; I_1 = i_1, \ldots, I_{t-1} = i_{t-1}\right] \\ &= \sum_{i_1} \cdots \sum_{i_{t-1}} \mathbf{Pr}\left[I_1 = i_1, \ldots, I_{t-1} = i_{t-1}\right] \ell_i^{(t)} = \ell_i^{(t)} \enspace .\end{aligned}$$

Furthermore, $\sum_{t=1}^T \sum_{i=1}^n \mathbf{E}\left[\tilde{\ell}_i^{(t)}\right] = \sum_{t=1}^T \sum_{i=1}^n \ell_i^{(t)} \leq nT$.

For the term $\mathbf{E}\left[q_i^{(t)} \tilde{\ell}_i^{(t)}\right]$ on the left-hand side, we have to be a bit more careful because both $q_i^{(t)}$ and $\tilde{\ell}_i^{(t)}$ are random variables, which are correlated in a complicated way. (We defined $\tilde{\ell}_i^{(t)}$ based on $q_i^{(t)}$.) Again, we fix $I_1, \ldots, I_{t-1}$ arbitrarily and, this way, $q_i^{(t)}$ is not random anymore. So, we now get

$$\mathbf{E}\left[q_i^{(t)} \tilde{\ell}_i^{(t)} \;\middle|\; I_1, \ldots, I_{t-1}\right] = q_i^{(t)} \mathbf{E}\left[\tilde{\ell}_i^{(t)} \;\middle|\; I_1, \ldots, I_{t-1}\right] = q_i^{(t)} \ell_i^{(t)} \enspace .$$

Now, take the expectation over $I_1, \ldots, I_{t-1}$. Fortunately, $\ell_i^{(t)}$ is not random, therefore

$$\mathbf{E}\left[q_i^{(t)} \ell_i^{(t)}\right] = \mathbf{E}\left[q_i^{(t)}\right] \ell_i^{(t)} = \mathbf{Pr}\left[I_t = i\right] \ell_i^{(t)} \enspace .$$

So, we also have

$$\sum_{t=1}^T \sum_{i=1}^n \mathbf{E}\left[q_i^{(t)} \tilde{\ell}_i^{(t)}\right] = \sum_{t=1}^T \sum_{i=1}^n \mathbf{Pr}\left[I_t = i\right] \ell_i^{(t)} = \mathbf{E}\left[\sum_{t=1}^T \ell_{I_t}^{(t)}\right] \enspace .$$

□

The bound in Theorem 20.1 depends on γ. Note that γ can be thought of balancing off *exploration* and *exploitation*. If we set γ to 0, then once an action has turned out to be bad it will rarely be chosen in the future because it is always reported to have high cost. If we set γ to 1, then we ignore the history when making our decision. The parameter γ has to be chosen carefully so that actions still have a chance to recover (meaning that we explore) but we keep choosing the actions that turned out to be good so far.

If we set $\gamma = \eta = \sqrt[3]{\frac{n \ln n}{T}}$, then Theorem 20.1 gives us

$$\mathbf{E}\left[\sum_{t=1}^T \ell_{I_t}^{(t)}\right] \leq \min_i \sum_{t=1}^T \ell_i^{(t)} + \frac{n \ln n}{\gamma \eta} + (\eta + \gamma) T = \min_i \sum_{t=1}^T \ell_i^{(t)} + 3 (n \ln n)^{1/3} T^{2/3} \enspace .$$

So the regret is bounded by $3(n \ln n)^{1/3} T^{2/3}$. As a matter of fact, the same algorithm with different choice of η and γ and only a more careful, but more complex analysis also gives a regret bound of $O(\sqrt{Tn \log n})$. Remember that for the experts setting, the bound was $O(\sqrt{T \log n})$.