Algorithms and Uncertainty, Winter 2017/18 Lecture 9 (5 pages)

No-Regret Learning: Multi-Armed Bandits

Instructor: Thomas Kesselheim

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 an randomized way), which we call I_t. Then the algorithm gets to know the entire vector $\ell^{(t)}$.

If $\ell_i^{(t)} \in [0, \rho]$ we showed that Randomized Weighted Majority (RWM) is a randomized algorithm (with parameter η) that guarantees

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

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

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

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

An algorithm that guarantees $R^{(T)} = o(T)$ 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.

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, that is the algorithm should ensure that for all sequences $\ell^{(1)}, \ldots, \ell^{(T)}$, the regret

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

grows sublinearly, that is, $R^{(T)} = o(T)$.

3 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. It is important to note here that this is not the optimal algorithm and analysis in terms of the regret. However, the optimal one uses exactly these ideas and is just a little more careful but more difficult.

The idea is as follows: We run an experts algorithm like RWM 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 $\ell_{I_t}^{(t)}$, $\ell_i^{(t)}$ for $i \neq I_t$ is 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: $p_i^{(t)}$ can be arbitrarily small, so $\tilde{\ell}_i^{(t)}$ is unbounded. Our algorithm, however, only 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.

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.

Theorem 9.1. *When using RWM 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 $\tilde{\ell}^{(1)}, \ldots, \tilde{\ell}^{(n)}$ that is given to RWM. What would RWM do on this sequence? It computes probability fectors $p^{(1)}, \ldots, p^{(T)}$. 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 RWM 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

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

This inequality still talks about the fake sequence $\tilde{\ell}^{(1)}, \ldots, \tilde{\ell}^{(T)}$ but we 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 $I_1, \ldots, I_{t-1}$ arbitrarily. This fixes $q^{(t)}$ and $\mathbf{Pr}\left[I_t = i \mid I_1, \ldots, I_{t-1}\right] = q_i^{(t)}$. So

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

for *any* choices of $I_1, \ldots, I_{t-1}$. So, also $\mathbf{E}\left[\tilde{\ell}_i^{(t)}\right] = \ell_i^{(t)}$. 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. We again fix $I_1, \ldots, I_{t-1}$ arbitrarily and this way, $q^{(t)}$ is not random anymore. So, we now get

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

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

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

So, we also have

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

□

The bound in Theorem 9.1 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 9.1 gives us

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

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 $O(\sqrt{T \log n})$.

4 Lower Bound on the Regret

Theorem 9.2. *Even for $n = 2$, no algorithm guarantees external regret $o(\sqrt{T})$.*

Proof. Let T be an even square number. We generate a random sequence $\ell^{(1)}, \ldots, \ell^{(T)}$. For each t, we set $\ell^{(t)}$ independently to $(1, 0)$ or to $(0, 1)$ with probability $1/2$ each. Observe that in each step, no matter how the algorithm chooses the probabilities, its expected cost will be $1/2$. So $\mathbf{E}\left[L_{\mathrm{Alg}}^{(T)}\right] = T/2$, where the expectation is also over the randomization of the sequence.

We have to compare this to $\mathbf{E}\left[\min_i L_i^{(T)}\right]$. We will show that $\mathbf{E}\left[\min_i L_i^{(T)}\right] = T/2 - \Omega(\sqrt{T})$. Note that $L_1^{(T)}$ and $L_2^{(T)}$ are identically distributed, namely according to a binomial distribution with parameters T and $1/2$. So they are the number of times we see heads in T independent fair coin tosses.

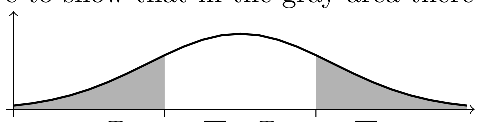
Furthermore, $L_1^{(T)} + L_2^{(T)} = T$. So, $\min_i L_i^{(T)}$ never exceeds $T/2$. Therefore, we can write

$$\begin{aligned}\mathbf{E}\left[\min_i L_i^{(T)}\right] &\leq \mathbf{Pr}\left[\min_i L_i^{(T)} < \frac{T}{2} - \alpha\sqrt{T}\right]\left(\frac{T}{2} - \alpha\sqrt{T}\right) + \mathbf{Pr}\left[\min_i L_i^{(T)} \geq \frac{T}{2} - \alpha\sqrt{T}\right]\frac{T}{2} \\ &\leq \frac{T}{2} - \alpha\sqrt{T} + \alpha\sqrt{T}\mathbf{Pr}\left[\min_i L_i^{(T)} \geq \frac{T}{2} - \alpha\sqrt{T}\right] \quad .\end{aligned}$$

We have $\min_i L_i^{(T)} \geq \frac{T}{2} - \alpha\sqrt{T}$ if and only if $\frac{T}{2} - \alpha\sqrt{T} \leq L_1^{(T)} \leq \frac{T}{2} + \alpha\sqrt{T}$, so

$$\mathbf{Pr}\left[\min_i L_i^{(T)} \geq \frac{T}{2} - \alpha\sqrt{T}\right] = \mathbf{Pr}\left[\frac{T}{2} - \alpha\sqrt{T} \leq L_1^{(T)} \leq \frac{T}{2} + \alpha\sqrt{T}\right] \quad .$$

We have to show that $L_1^{(T)}$ is not always close to its expectation (which is $T/2$). Pictorially, we have to show that in the gray area there is at least a constant probability.



As $L_1^{(T)}$ is binomially distributed, we have

$$\mathbf{Pr}\left[\frac{T}{2} - \alpha\sqrt{T} \leq L_1^{(T)} \leq \frac{T}{2} + \alpha\sqrt{T}\right] = \sum_{j=\frac{T}{2}-\alpha\sqrt{T}}^{\frac{T}{2}+\alpha\sqrt{T}} \mathbf{Pr}\left[L_1^{(T)} = j\right] \text{ and } \mathbf{Pr}\left[L_1^{(T)} = j\right] = \frac{1}{2^T}\binom{T}{j} \quad .$$

We have to bound the binomial coefficient. We can do this using Stirling's approximation, which says $\sqrt{2\pi}\ k^{k+\frac{1}{2}}\mathrm{e}^{-k} \leq k! \leq \mathrm{e}\ k^{n+\frac{1}{2}}\mathrm{e}^{-k}$ for all k. This gives us $\binom{T}{T/2} \leq \frac{\mathrm{e}}{\pi}\frac{2^T}{\sqrt{T}}$. Using the monotonicity of binomial coefficients, we have $\binom{T}{j} \leq \frac{\mathrm{e}}{\pi}\frac{2^T}{\sqrt{T}}$ for all j. So

$$\mathbf{Pr}\left[L_1^{(T)} = j\right] = \frac{1}{2^T}\binom{T}{j} \leq \frac{\mathrm{e}}{\pi}\frac{1}{\sqrt{T}}$$

and therefore

$$\mathbf{Pr}\left[\min_i L_i^{(T)} \geq \frac{T}{2} - \alpha\sqrt{T}\right] \leq 2\alpha\sqrt{T} \cdot \frac{\mathrm{e}}{\pi}\frac{1}{\sqrt{T}} = \frac{2\alpha\mathrm{e}}{\pi}$$

and also

$$\mathbf{E}\left[\min_i L_i^{(T)}\right] \leq \frac{T}{2} - \alpha\sqrt{T} + \alpha\sqrt{T}\frac{2\alpha\mathrm{e}}{\pi} \quad .$$

Using, for example, $\alpha = \frac{1}{2}$, we get $\mathbf{E}\left[\min_i L_i^{(T)}\right] \leq \frac{T}{2} - \frac{1}{2}(1 - \frac{\mathrm{e}}{\pi})\sqrt{T} \geq \frac{T}{2} - 0.06\sqrt{T}$. □

5 Unknown Time Horizon

So far, our algorithms assumed that we know the time horizon T. Indeed, with a slight modification, they also work for unknown time horizons.

The modified algorithm works as follows. Phase $k \geq 0$ consists of steps $2^k, \ldots, 2^{k+1} - 1$. So it consists of 2^k steps. At the beginning of a phase, we restart the no-regret algorithm with $T' = 2^k$.

Let us analyze RWM in this construction.

Theorem 9.3. *The modified version of RWM has regret $O(\sqrt{T \log n})$.*

Proof. We start $m = \lfloor \log_2 T \rfloor + 1$ phases during T steps. As the last phase might not be complete, we fill up the sequence by $\ell^{(T+1)}, \ldots, \ell^{(2^m-1)}$ with all-zero vectors. This neither changes the cost of a single action nor of the algorithm.

In each phase, we restart the algorithm. Therefore, if P_k are the steps in phase k, we have the regret guarantee

$$\sum_{t \in P_k} \sum_{i=1}^{n} p_i^{(t)} \ell_i^{(t)} \leq \min_i \sum_{t \in P_k} \ell_i^{(t)} + 2\sqrt{|P_k| \ln n} \enspace .$$

Now, we take the sum over $k = 0, \ldots, m$ on both sides

$$\sum_{k=0}^{m-1} \sum_{t \in P_k} \sum_{i=1}^{n} p_i^{(t)} \ell_i^{(t)} \leq \sum_{k=0}^{m-1} \min_i \sum_{t \in P_k} \ell_i^{(t)} + 2 \sum_{k=0}^{m-1} \sqrt{|P_k| \ln n}$$

The first sum, $\sum_{k=0}^{m} \sum_{t \in P_k} \sum_{i=1} n p_i^{(t)} \ell_i^{(t)}$ is exactly the cost of the algorithm.

For the second sum, we have

$$\sum_{k=0}^{m-1} \min_i \sum_{t \in P_k} \ell_i^{(t)} \leq \min_i \sum_{k=0}^{m-1} \sum_{t \in P_k} \ell_i^{(t)} = L_i^{(T)} \enspace .$$

And for the third sum, we use that $|P_k| = 2^k$, which gives

$$2 \sum_{k=0}^{m-1} \sqrt{|P_k| \ln n} = 2\sqrt{\ln n} \sum_{k=0}^{m-1} (\sqrt{2})^k = 2\sqrt{\ln n} \frac{(\sqrt{2})^m - 1}{\sqrt{2} - 1} = O(\sqrt{T \ln n}) \enspace .$$

$\square$