Online Computation, Armed Bandit.

David Ponarovsky

March 11, 2023

multi armed bandits

1. n arms.
2. $l_{i,t}$ and $g_{i,t}$ are the lost and the gain function of the ith arm at time t.
3. In each time step t we learn only the tth coordinate of the arm we chooses.

When we choose i_t we only learn l_i, t, unlikely of the experts case when we learn the loss of all the experts, here we only learn the loss of the arm that we choose, we don't know what is the loss of the others arms. So it seems like hopeless situation, and deterministic that indeed the case, but it turns out that using randomness we can actually do something.

So, we will aim at competing against adaptive adversary, But unfortunately here it is more complicated situation, so we will discuss at first an oblivious adversary. So what is mean that our adversary is oblivious in that case? The adversary could choose the loss vector depends on the our last chooses in past, but he doesn't know what is going be our current choose. So if our choices are random variables than the loss function is also going to be a random variable.

We measure the performance by a regret function define by:

$$R_T = \sum_{t=1}^{T} l_{i_t,t} - \min \sum_{t=1}^{T} l_{j,t}$$

When T count the number of steps. The goal is to bound the $\mathbf{E}[R_T]$. Notice that algorithm hasn't the information to calculate that quantity. In general good algorithms combine two consideration. So first at all they want to exploit arms that have low cost, But if we just exploit arms that seems to be promise in the begging we might ends with choice that has an heigh cost. So there is must to be an element of exploration, and this is the secret. Blanching right between exploration and exploition.

Another representation of the Regert is the difference between the score we gained and the gained which achieved by gamble on the best bandit.

$$R_T = \max \sum_{t=1}^{T} g_{j,t} - \sum_{t=1}^{T} g_{i_t,t}$$

Let's define the pseudo-regret,

$$\bar{R}_T = \mathbf{E}\left[\sum_{t=1}^{T} l_{i_t,t}\right] - \mathbf{E}\left[\min \sum_{t=1}^{T} l_{j,t}\right] \leq \mathbf{E}[R_T]$$

So what we are planning to do, is first to design an algorithm that has a good pseudo-regret gurntes. But that doesn't says anything about the regret gurntess. But that will be our starting point, and when the adversary is a oblivious the pseudo regret and the regret are equivalence. So one way to think about that is that we looking for way to bound the regret against the oblivious adversary and we think how we can improve the algorithm against a non oblivious adversary.

We will show that, at least amortized, the regret per step will be a little $o(1)$. In the experts case the regret after T steps were something like $\sqrt{T \log(n)}$. Here the factor is going to be something like $O(n \log n)$. This is not surprising, because in the experts case we learn all the value, while here we learn only a single bit of information, So it make sense that we will have to pay a linear factor of time till the converge into the stage in which the regret is arbitrary small.

The first algorithm is called **EXP3**.

So how that algorithm is operating? First we initialize for any $i \in \{1, .., n\}$, $\tilde{L}_{i,0} \leftarrow 0$. We are going to mimic the multiplicative update or Hedge in the experts case, the problem is that we don't know the values. Because, the update in the experts case require to know the loss of all the experts. To overcome over that, we will try to estimate the loss of the other by unbias estimators, or learn the L's function by the way. In addition, we initialize a distribution over the choosing the arms P_t which in time zero is just the uniform distribution over the arms.

```
for i ∈ [n] set L̃_{i,0} ← 0
set P_t to the uniform distribution on {1, 2..n}
for t ∈ 1, 2.., T do
    draw t_t from P_t
    for i ∈ [n] do
        L̃_{i,t} ← L̃_{i,t−1} + 1_{i_t=i} · l_{i,t} / P_{i,t}
    end
    update the probability P_{i,t+1} ← e^{−η_t L̃_{i,t}} / Σ_j e^{−η_t L̃_{i,t}}
    the η_t are non-increasing sequence
     of values in (0, ∞) that will determined later.
end
```

The ηt is called the learning rate. We will see two different versions, the first with fixed learning rate and the other about choosing a learning rate that goes down.

Claim. $\tilde{l}_{i,t}$ is unbiased estimator of$l_{i,t}$.

Proof. Let's proof that, we just have to show that the expectation of that value over the possibilities is just $l_{i,t}$.

$\mathbf{E}\left[\tilde{l}_{i,t}\right] = \frac{l_{i,t}}{P_{i,t}} \cdot \mathbf{Pr}\left[i_t = i\right] = l_{i,t}$. When in the last pass we use the fact that the algorithm draw from P_t.

Claim. For every non-increasing $\{\eta\}$, EXP3 satisfies the following ineqaulity:

$$\bar{R}_T \leq \frac{n}{2}\sum_{t=1}^{T}\eta_t + \frac{\ln(n)}{\eta_T}$$

Corollary. We can set the following two bounds.

1. For $\eta_t = \sqrt{\frac{\ln(n)}{nT}}$ we have that $\bar{R}_T \leq \sqrt{2Tn\ln(n)}$.
2. For $\eta_t = \sqrt{\frac{\ln(n)}{nt}}$ we have that $\bar{R}_T \leq 2\sqrt{Tn\ln(n)}$ (expand by factor of $\sqrt{2}$).

The disadvantage of the first bound, is that using it require us knowing what is the number of the steps for initialized the rate. This result tell us that if we compete against oblivious adversary after $n\log(n)$ steps the regret increase by $o(1)$.

Proof. Recall that $\tilde{l}_{i,t}$ is an unbiased estimator of l and notice that $\mathbf{E}\left[\left(\tilde{l}_{i,t}\right)^2\right] = \frac{l_{i,t}^2}{P_{i,t}}$ and that $\mathbf{E}\left[\frac{1}{p_{i,t}}\right] = n$. Lets look on the sum $\sum_{t=1}^{T} l_{i_t,t} - \sum_{t=1}^{T} l_{j,t}$

$$\sum_{t=1}^{T} l_{i_t,t} - \sum_{t=1}^{T} l_{j,t} = \sum_{t=1}^{T}\mathbf{E}\left[\tilde{l}_{i_t,t}\right] - \sum_{t=1}^{T}\mathbf{E}\left[\tilde{l}_{j,t}\right]$$

$$\mathbf{E}\left[\tilde{l}_{i_t,t}\right] = \frac{1}{\eta_t}\ln\left(e^{\eta_t\mathbf{E}\left[\tilde{l}_{i,t}\right]}\right) + \frac{1}{\eta_t}\ln\mathbf{E}\left[e^{-\eta_t\tilde{l}_{i,t}}\right] - \frac{1}{\eta_t}\ln\mathbf{E}\left[e^{-\eta_t\tilde{l}_{i,t}}\right]$$

$$= \frac{1}{\eta_t}\ln\mathbf{E}\left[e^{\eta_t\mathbf{E}\left[\tilde{l}_{i,t}\right]-\eta_t\tilde{l}_{i,t}}\right] - \frac{1}{\eta_t}\ln\mathbf{E}\left[e^{-\eta_t\tilde{l}_{i,t}}\right]$$

$$=$$