AI-3001: Advanced Topics in Machine Learning **Aug-Nov 2022**

Lecture 12: Bandit Optimization

Lecturer: Ganesh Ghalme *Scribes: Ganesh Ghalme*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

12.1 Bandits Setting

Algorithm 1: Multi-armed Bandits (MAB) setting

```
Input: Number of arms $n$
Initialize: Environment selects the loss vectors $\{\ell_t\}_{t\geq 1}$ a-priori.
for $t = 1, 2, \cdots$ do
    - Algorithm picks (pulls) an arm $i_t \in [n]$;
    - Environment reveals the loss $\ell_{i_t,t} \in [0, 1]$ to the algorithm
end
```

- The environment sets the loss functions $(\ell_t)_{t\geq 1}$ beforehand and reveals only the loss corresponding to i_t i.e. the arm pulled at time t to the algorithm.

- Algorithm is blind to rewards for arms it did not pull; hence must work with severely limited information

- Exploration-Exploitation dilemma: Should an algorithm select an arm that has least loss in the past (from history) vs should an algorithm pull a random arm to learn their losses

- Throughout we will consider that the losses are bounded in $[0, 1]$, n is finite and stopping time T is unknown beforehand

- We consider randomized algorithms i.e. $i_t \sim \mathcal{D}_t$ for some distribution $\mathcal{D}_t \in \Delta_n$. However, for the first part of this module (adversarial bandits) we will consider that the losses are given by the environment in adversarial manner

- Why not $\ell_t = 0$ for all t? Because its regret is 0 and as an adversary the environment can do better...

12.2 Naive application of OGD

```
Algorithm 2: OGD for MAB
Input: Number of arms $n$, $\delta \in (0,1)$
for $t = 1, 2, \cdots$ do
    $b_t \sim Bern(\delta)$ ;
    if $b_t = 1$ then
        - $i_t \sim Unif([n])$ Exploration ;
        - $\widehat{\ell}_{i,t} = \begin{cases} \frac{n}{\delta}\ell_{i,t} & \text{if } i = i_t \\ 0 & \text{otherwise} \end{cases}$ ;
        - Set $\widehat{f}_{i,t} = \widehat{\ell}_{i,t} \cdot x_{i,t}$ for all $i \in [n]$. i.e. $\widehat{f}_t = \langle \widehat{\ell}_t, x \rangle$;
        - $x_{t+1} = OGD(\widehat{f}_1, \widehat{f}_2, \cdots \widehat{f}_t)$;
    end
    else if $b_t = 0$ then
        - $i_t \sim x_t$ Exploitation;
        - $x_{t+1} = x_t$, $\widehat{\ell}_t = 0$ and $\widehat{f}_t = 0$;
    end
end
```

Lemma 12.1. $\mathbb{E}[\ell_{i_t,t}] \leq \mathbb{E}[\langle \widehat{\ell}_t, x_t \rangle] + \delta$

Proof.

$$\begin{aligned} \mathbb{E}[\ell_{i_t,t}] &= \mathbb{P}(b_t = 1)\mathbb{E}[\ell_{i_t,t}|b_t = 1] + \mathbb{P}(b_t = 0)\mathbb{E}[\ell_{i_t,t}|b_t = 0] \\ &\leq \delta + (1-\delta)\mathbb{E}[\ell_{i_t,t}|b_t = 0] && \text{(since } \mathbb{E}[\ell_{i_t,t}|b_t = 1] \leq 1) \\ &= \delta + (1-\delta)\mathbb{E}[\langle \ell_t, x_t \rangle | b_t = 0] && (\ b_t = 0 \text{ implies } i_t \sim x_t) \end{aligned}$$

We note that for $\langle \ell_t, x_t \rangle$ is a non-negative r.v. hence we use the result $\mathbb{E}[X] \geq \mathbb{E}[X, E]$ for any event E (proof left as an exercise).

$$\begin{aligned} &\leq \delta + \mathbb{E}[\langle \ell_t, x_t \rangle] \\ &= \delta + \mathbb{E}[\langle \widehat{\ell}_t, x_t \rangle] && (\text{ Since } \widehat{\ell}_t \text{ is an unbiased estimator of } \ell_t) \end{aligned}$$

□

Using the above lemma we now prove the regret bound.

Theorem 12.2. *The regret of the $OGD - MAB$ is upper bounded by*

$$\mathbb{E}[\mathcal{R}_T(OGD - MAB)] \leq 4n^{2/3}T^{2/3} \tag{12.1}$$

Proof.

$$
\begin{aligned}
\mathbb{E}[\mathcal{R}_T(OGD - MAB)] &= \mathbb{E}\Big[\sum_{t=1}^{T}\ell_{i_t,t} - \sum_{t=1}^{T}\ell_{i^\star,t}\Big] \\
&= \mathbb{E}\Big[\sum_{t=1}^{T}\ell_{i_t,t} - \sum_{t=1}^{T}\widehat{\ell}_{i^\star,t}\Big] && \text{(Since $\widehat{\ell}$ is an unbiased estimator)} \\
&\leq \mathbb{E}\Big[\sum_{t=1}^{T}\langle \ell_t, x_t\rangle - \sum_{t=1}^{T}\widehat{\ell}_{i^\star,t}\Big] + \delta T && \text{(From above lemma)} \\
&= \mathbb{E}[\mathcal{R}_T(OGD)] + \delta T \\
&\leq \frac{3}{2}GD\sqrt{T} + \delta T
\end{aligned}
$$

Here $G = \sup||\widehat{\ell}|| \leq \frac{n}{\delta}$ and $D \leq 2$ (proof left as exercise). We have

$$
\begin{aligned}
&\leq \frac{3n}{\delta}\sqrt{\delta T} + \delta T \\
&\leq \frac{3n}{\sqrt{\delta}}\sqrt{n} + \delta T \\
&= 3n^{2/3}T^{1/3} + n^{2/3}T^{2/3} && \text{(Choosing $\delta = n^{2/3}T^{-1/3}$)} \\
&\leq 4n^{2/3}T^{2/3}
\end{aligned}
$$

□

Some features of the above algorithm.

- It is exploration separated (either does exploration or exploitation in each round).
- The estimates for the loss function is updated only in the exploration phase. Hence new information is not obtained in $(1-\delta)T = T - n^{2/3}T^{2/3}$ expected number of rounds. This leads to increased regret.
- Next we will show that one can do better by simulteneously exploring and exploiting in the same round. In fact, one can do as good as (in order terms) the full information setting.

12.3 EXP3-γ

```
Algorithm 3: EXP3-$\gamma$
Input: Number of arms $n$, $\gamma \in (0,1)$
Initialize: $w_{i,1} = 1$ for all $i \in [n]$
for $t = 1, 2, \cdots$ do
    - $p_{i,t} = (1-\gamma)\frac{w_{i,t}}{\sum_{i=1}^{n} w_{i,t}} + \gamma\frac{1}{n}$ ;
    - Draw $i_t \sim p_t$ where $p_t = \begin{pmatrix} p_{1,t} \\ p_{2,t} \\ . \\ . \\ . \\ p_{n,t} \end{pmatrix}$
    - Observe $r_{i_t,t}$, the reward at time $t$ (from arm $i_t$);
    - $\widehat{r}_{i,t} = \begin{cases} \frac{r_{j,t}}{p_{j,t}} & \text{if } i_t = i \\ 0 & \text{otherwise} \end{cases}$ ;
    - $w_{i,t+1} = w_{i,t} \cdot e^{\frac{\gamma}{n}\widehat{r}_{i,t}}$ ;
end
```

Theorem 12.3. *For any $\gamma \in (0,1)$, any reward sequence $(r_t)_{t\geq 1}$ with $r_t \in [0,1]^n$ we have*

$$\max_{i=1,2,\cdots,n} \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{i_t,t} \leq 2\gamma T + \frac{n\log(n)}{\gamma} \tag{12.2}$$

Proof. We will follow the potential argument we used for the proof of exponential weights algorithm. Let $W_t = \sum_{i=1}^{n} w_{i,t}$. Notice here that it is the sum of current weights not previous weights (it shouldn't be a problem, the argument remains essentially the same). For the upper bound we have

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \sum_{i=1}^{n} \frac{w_{i,t+1}}{W_t} = \sum_{i=1}^{n} \left(\frac{w_{i,t}}{W_t}\right) e^{\gamma \widehat{r}_{i,t}/n} \\ &= \sum_{i=1}^{n} \frac{p_{i,t} - \gamma/n}{1-\gamma} \cdot exp(\frac{\gamma \widehat{r}_{i,t}}{n}) \\ &\leq \sum_{i=1}^{n} \frac{p_{i,t} - \gamma/n}{1-\gamma} \cdot exp(\frac{\gamma \widehat{r}_{i,t}}{n}) \end{aligned}$$

Notice $\frac{\gamma \widehat{r}_{i,t}}{n} \leq 1$ for all i (convince yourself!)

$$\begin{aligned} &\leq \sum_{i=1}^{n} \frac{p_{i,t} - \gamma/n}{1-\gamma} (1 + \frac{\gamma}{n}\widehat{r}_{i,t} + \left(\frac{\gamma}{n}\right)^2 \widehat{r}_{i,t}^2) \qquad \text{(Using } e^x \leq 1 + x + x^2 \text{ for all } x \in [0,1]) \\ &\leq \underbrace{\sum_{i=1}^{n} \frac{p_{i,t} - \gamma/n}{1-\gamma}}_{=1} + \frac{\gamma/n}{1-\gamma} \underbrace{\sum_{i=1}^{n} p_{i,t}\widehat{r}_{i,t}}_{=r_{i_t,t}} + \frac{(\gamma/n)^2}{1-\gamma} \sum_{i=1}^{n} p_{i,t}\widehat{r}_{i,t}^2 \\ &= 1 + \frac{\gamma/n}{1-\gamma} r_{i_t,t} + \frac{(\gamma/n)^2}{1-\gamma} \sum_{i=1}^{n} \widehat{r}_{i,t} \end{aligned}$$

To understand the last term notice that $\sum_{i=1}^{n} p_{i,t}\widehat{r}_{i,t}^{2} = \sum_{i=1}^{n} \underbrace{(p_{i,t}\widehat{r}_{i,t})}_{\leq 1} \widehat{r}_{i,t} \leq \sum_{i=1}^{n} 1 \cdot \widehat{r}_{i,t}$. The last inequality follows from the fact that all $\widehat{r}_{i,t}$ are non-negative. Now taking log both sides and using the inequality $\log(1+x) \leq x$ we have

$$\log\left(\frac{W_{t+1}}{W_t}\right) \leq \frac{\gamma/n}{1-\gamma} r_{i_t,t} + \frac{(\gamma/n)^2}{1-\gamma} \sum_{i=1}^{n} \widehat{r}_{i,t}$$

$$\implies \log\left(\frac{W_{T+1}}{W_1}\right) \leq \frac{\gamma/n}{1-\gamma} \sum_{t=1}^{T} r_{i_t,t} + \frac{(\gamma/n)^2}{1-\gamma} \sum_{t=1}^{T}\sum_{i=1}^{n} \widehat{r}_{i,t} \qquad \text{(telescoping sum over } T \text{ rounds)}$$

Next we obtain the lower bound over $\log\left(\frac{W_{T+1}}{W_1}\right)$

$$\begin{aligned}\log\left(\frac{W_{T+1}}{W_1}\right) &\geq \log\left(\frac{w_{T+1}}{W_1}\right) \\ &= \log(e^{\gamma/n \sum_{t=1}^{T} \widehat{r}_{i,t}}) - \log(n) \qquad \text{(as } W_1 = n\text{)} \\ &= \gamma/n \sum_{t=1}^{T} \widehat{r}_{i,t} - \log(n)\end{aligned}$$

Simplifying

$$\sum_{t=1}^{T} \widehat{r}_{i,t} - \frac{n}{\gamma}\log(n) \leq \frac{1}{1-\gamma}\sum_{t=1}^{T} r_{i_t,t} + \frac{\gamma/n}{1-\gamma}\sum_{t=1}^{T}\sum_{i=1}^{n} \widehat{r}_{i,t}$$

$$\iff \sum_{t=1}^{T} \widehat{r}_{i,t} - \sum_{t=1}^{T} r_{i_t,t} \leq \gamma \sum_{t=1}^{T} \widehat{r}_{i,t} + \frac{(1-\gamma)}{\gamma} n\log(n) + \gamma/n \sum_{i=1}^{n}\sum_{t=1}^{T} \widehat{r}_{i,t}$$

Taking expectation both sides and letting $i^* = \arg\min_i \sum_{t=1}^{T} r_{i,t}$, we have

$$\begin{aligned}&\mathbb{E}[\mathcal{R}_T(EXP3-\gamma)] \leq \frac{n\log(n)}{\gamma} + 2\gamma \sum_{t=1}^{T} r_{i^*,t} \\ &\leq \frac{n\log(n)}{\gamma} + 2\gamma T\end{aligned}$$

□