E1 245: Online Prediction & Learning **Fall 2014**

Lecture 13 — September 16

Lecturer: Aditya Gopalan *Scribe: Sayak Ray Chowdhury*

13.1 Recall

In the previous lecture, we have seen EXP3 algorithm regarding bandit model of learning i.e online learning with partial information. The original work by Auer et. al. [1] considers a slightly different version of what we have seen. Here we present the algorithm given in [1] and denote it by EXP3-ORIG.

EXP3-ORIG

Parameter: $\eta \in [0,1]$

Initialize: $p_1 = Uniform\{1,2,...,N\}$; $\tilde{L_0} = \mathbf{0} \in \mathbb{R}^N$

At each time t = 1, 2, 3,...,T

1. Sample $I_t \sim p_t$, where $p_t \equiv (p_{i,t})_{i=1}^N$

For each i = 1, 2,..., N

2. $\tilde{l_{i,t}} := \frac{l_{i,t}}{p_{i,t}} \mathbb{1}\{I_t = i\}$, where $l_t \equiv (l_{i,t})_{i=1}^N$, $\tilde{l_t} \equiv (\tilde{l_{i,t}})_{i=1}^N$

3. $\tilde{L_{i,t}} := \tilde{L_{i,t-1}} + \tilde{l_{i,t}}$, where $\tilde{L_t} \equiv (\tilde{L_{i,t}})_{i=1}^N$

4. $p_{i,t+1} := \dfrac{(1-\eta)\exp(-\eta \tilde{L_{i,t}})}{\sum_{j=1}^{N} \exp(-\eta \tilde{L_{j,t}})} + \dfrac{\eta}{N}$

From the previous lecture, we know $\mathbb{E}[Regret_T^{EXP3}] \le O(\sqrt{NT\log N})$. This **optimal** bound holds also for EXP3-ORIG. In this lecture, we will :

– Show an EXP3 like algorithm that enjoys regret bound with high probability (WHP).
– Give a lower bound on regret (MINIMAX regret) across all bandit algorithms.

13.2 Modification of EXP3 to get WHP regret

To motivate this modification, first we argue about a technical issue (though very loose) with EXP3-ORIG and equivalently with EXP3 : Variance of estimated losses $\tilde{L_{i,t}}$, which are unbiased estimates of $L_{i,t}$ can be very large for both the algorithms. First, lets see why it is so.
Recall the definition of $\tilde{l_{i,t}}$. We have,

$\mathbb{E}[\tilde{l_{i,t}}^2 \mid \mathscr{F}_{t-1}] = p_{i,t} \dfrac{{l_{i,t}}^2}{{p_{i,t}}^2} = \dfrac{{l_{i,t}}^2}{p_{i,t}} \approx O(1/p_{i,t})$

Now, in EXP3-ORIG, optimal $\eta \approx \frac{1}{\sqrt{T}}$ and $p_{i,t} \geq \frac{\eta}{N} \approx \frac{1}{N\sqrt{T}}$
So we have, $\mathrm{Var}[\tilde{l_{i,t}} \mid \mathscr{F}_{t-1}] \approx O(\sqrt{T})$
Hence, $\mathrm{Var}[\tilde{L_{i,t}} \mid \mathscr{F}_{t-1}] \approx O(T^{3/2})$, which gets very large as T becomes larger.
Similarly, for EXP3 also $p_{i,t}$ can be very small, thus making the variance very large. To overcome this issue, we make two key tweaks in EXP3-ORIG :
(i) Let us consider **rewards** or gains instead of losses, i.e.
gains $g_{i,t} := 1 - l_{i,t}$; $g_{i,t} \in [0,1]$ and
gain estimates $\tilde{g_{i,t}} := \frac{g_{i,t}}{p_{i,t}} \mathbb{1}\{I_t = i\}$. Note, $\tilde{g_{i,t}} \neq 1 - \tilde{l_{i,t}}$
(ii) Control variance of (gain) estimates by adding a stabilization term (β) :

$$
\begin{aligned}
g'_{i,t} &:= \frac{g_{i,t} + \beta}{p_{i,t}}, \text{ if } I_t = i \\
&:= \frac{\beta}{p_{i,t}}, \text{ if } I_t \neq i \\
&= \tilde{g_{i,t}} + \frac{\beta}{p_{i,t}}
\end{aligned}
$$

The underlying idea behind these tweaks is to ensure that $G'_{i,t}$ is an **upper confidence bound** for $G_{i,t}$, where $G_{i,t} = \sum_{s=1}^{t} g_{i,s}$ and $G'_{i,t} = \sum_{s=1}^{t} g'_{i,s}$
Now, we present the WHP version of the EXP3-ORIG algorithm and call it as EXP3.P as given in [1].

EXP3.P

Parameters: $\beta, \gamma, \eta \in [0,1]$
Initialize: $p_1 = Uniform\{1,2,...,N\}$; $G'_{i,0} = 0$, $\forall i \in [N]$
At each time t = 1, 2, 3,...,T
1. Sample $I_t \sim p_t$
For each i = 1, 2,..., N
2. $g'_{i,t} := \frac{g_{i,t}\, \mathbb{1}\{I_t = i\} + \beta}{p_{i,t}}$
3. $G'_{i,t} := G'_{i,t-1} + g'_{i,t}$
4. $p_{i,t+1} := \frac{(1-\gamma)\exp(\eta G'_{i,t})}{\sum_{j=1}^{N} \exp(\eta G'_{j,t})} + \frac{\gamma}{N}$

Theorem 13.1. *[Regret bound for EXP3.P]*
For any $\delta \in (0,1)$ with Probability $\geq (1-\delta)$, $Regret_T^{EXP3.P} \leq 5.15\sqrt{NT\log(N/\delta)}$
Now choosing δ to be small enough, the bound can be satisfied with high probability.

Proof: Before proving the theorem, first consider the following lemma :

Lemma 13.2. *[Upper Confidence property of g']*
For $\beta < 1$, with probability $\geq (1-\delta)$, $\sum_{t=1}^{T} g_{i,t} \leq \sum_{t=1}^{T} g'_{i,t} + \frac{\log(1/\delta)}{\beta}$, $\forall\ i \in [N]$

Proof: Let, $\mathscr{F}_{t-1} = \sigma - alg(I_1, I_2, ..., I_{t-1}, g(I_1, 1), g(I_2, 2) ..., g(I_{t-1}, t-1))$, where $g(i,t)$ denote $g_{i,t}$ and let $\mathbb{E}_t[.]$ denote $\mathbb{E}[.|\mathscr{F}_{t-1}]$. Now,
$\mathbb{E}_t[\exp(\beta g_{i,t} - \beta g'_{i,t})$
$= \mathbb{E}_t[\exp(\beta g_{i,t} - \frac{\beta g_{i,t}}{p_{i,t}} \mathbb{1}\{I_t = i\}) \exp(\frac{-\beta^2}{p_{i,t}})]$ [from definition of $g'_{i,t}$]
$= \mathbb{E}_t[\exp(\beta(g_{i,t} - \tilde{g_{i,t}})) \exp(\frac{-\beta^2}{p_{i,t}})]$ [from definition of $\tilde{g_{i,t}}$]
$= \mathbb{E}_t[1 + \beta(g_{i,t} - \tilde{g_{i,t}}) + (\beta(g_{i,t} - \tilde{g_{i,t}}))^2] \exp(\frac{-\beta^2}{p_{i,t}})$
[using, $e^x = 1 + x + x^2, \forall x \leq 1$, where $x = \beta(g_{i,t} - \tilde{g_{i,t}}) \leq \beta g_{i,t} \leq 1$ and using the fact that $p_{i,t}$ is measurable w.r.t $\mathscr{F}_{t-1}$]
$= (1 + \beta^2 \text{Var}_t[\tilde{g_{i,t}}]) \exp(\frac{-\beta^2}{p_{i,t}})$ [as, $\tilde{g}$ is an unbiased estimator of g]
$= (1 + \beta^2 \frac{g_{i,t}^2}{p_{i,t}}) \exp(\frac{-\beta^2}{p_{i,t}})$
$\leq \exp(\frac{\beta^2}{p_{i,t}}(g_{i,t}^2 - 1))$ [using, $1 + x \leq e^x, \forall x \geq 0$]
≤ 1
Multiplying over t = 1, 2,..., T we get, $\mathbb{E}[\exp(\beta G_{i,t} - \beta G'_{i,t})] \leq 1$
Now **Markov's inequality** gives,
$\text{Prob}[\beta G_{i,t} - \beta G'_{i,t} \geq \log(1/\delta)] \leq \frac{1}{1/\delta} = \delta$
Thus, with probability $\geq (1-\delta)$, $G_{i,t} \leq G'_{i,t} + \frac{\log(1/\delta)}{\beta}$
This is true for each i = 1, 2,..., N.

□

Now, it remains to prove theorem 13.1.
Proof of theorem: We will consider the **potential function** approach like exponential weights forecaster as discussed earlier. Let,
$W_t := \sum_{i=1}^{N} w_{i,t} := \sum_{i=1}^{N} \exp(\eta G'_{i,t})$.
So, $\log(\frac{W_T}{W_0}) = log(\frac{\sum_{i=1}^{N} \exp(\eta G'_{i,t})}{N}) \geq \eta \max_{i=1}^{N} G'_{i,t} - \log N$ ------- (*)
On the otherhand,
$\log(\frac{W_t}{W_{t-1}}) = \log(\sum_{i=1}^{N} (\frac{w_{i,t-1}}{W_{t-1}}) \exp(\eta g'_{i,t}))$

$$= \log(\sum_{i=1}^{N}(\frac{p_{i,t}-\gamma/N}{1-\gamma})\exp(\eta g'_{i,t})) \quad \text{[from definiton of EXP3.P update]}$$
$$= \log(\sum_{i=1}^{N}(\frac{p_{i,t}-\gamma/N}{1-\gamma})(1+\eta g'_{i,t}+\eta^2 g'^{\,2}_{i,t})) \quad \text{[setting } \eta \text{ s.t. } \eta g' \le 1 \text{ and using } e^x \le 1+x+x^2, \forall x \le 1]$$
$$\le log(1+\frac{\eta}{1-\gamma}\sum_i p_{i,t}\, g'_{i,t}+\frac{\eta^2}{1-\gamma}\sum_i p_{i,t}\, g'^{\,2}_{i,t})$$

Now, $\sum_i p_{i,t}\, g'^{\,2}_{i,t} = \sum_i p_{i,t}\, g'_{i,t}\,(\frac{g_{i,t}\,\mathbb{1}\{I_t=i\}+\beta}{p_{i,t}})$
$$= g'(I_t,t)g_{I_t,t}+\beta\sum_i g'_{i,t}$$
$$\le (1+\beta)\sum_i g'_{i,t}$$

And $\sum_i p_{i,t}\, g'_{i,t} = \sum_i p_{i,t}\,(\frac{g_{i,t}\,\mathbb{1}\{I_t=i\}+\beta}{p_{i,t}}) = g_{I_t,t}+N\beta$

Hence, $\log(\frac{W_t}{W_{t-1}}) \le \frac{\eta}{1-\gamma}(g_{I_t,t}+N\beta)+\frac{\eta^2}{1-\gamma}(1+\beta)\sum_i g'_{i,t}$ [as $\log(1+x) \le x, \forall x$]

Summing over t = 1, 2,..., T we get,

$log(\frac{W_T}{W_0}) \le \frac{\eta}{1-\gamma}G_T^{EXP3.P}+\frac{\eta N\beta T}{1-\gamma}+\frac{\eta^2}{1-\gamma}(1+\beta)\sum_{i=1}^{N} G'_{i,T}$ - - - - - - - (**)

where, $G_T^{EXP3.P} = \sum_{t=1}^{T} g_{I_t,t}$

Putting (*) and (**) together we get,

$\eta \max_i G'_{i,T} - logN \le \frac{\eta}{1-\gamma}G_T^{EXP3.P}+\frac{\eta N\beta T}{1-\gamma}+\frac{\eta^2}{1-\gamma}(1+\beta)\sum_{i=1}^{N} G'_{i,T}$

$\Rightarrow G_T^{EXP3.P} - (1-\gamma)G'_{max} \ge \frac{-(1-\gamma)}{\eta}logN - N\beta T - \eta(1+\beta)NG'_{max}$, where $G'_{max} = \max_i G_{i,t}$

$\Rightarrow G_T^{EXP3.P} \ge \frac{-logN}{\eta} - N\beta T + (1-\gamma-\eta(1+\beta)N)G'_{max}$ [as, $1-\gamma \le 1$]

Now, from lemma 13.2 we know,

w.p. $\ge (1-\delta)$, $G_{i,T} \le G'_{i,T}+\frac{\log(1/\delta)}{\beta}, \forall i \in [N]$.

Applying union bound,

w.p. $\ge (1-\delta)$, $\max_i G_{i,T} \le \max_i G'_{i,T}+\frac{\log(N/\delta)}{\beta}$

Hence w.p. $\ge (1-\delta)$,

$G_T^{EXP3.P} \ge -\frac{\log N}{\eta} - N\beta T + (1-\gamma-\eta(1+\beta)N)\max_i G_{i,T} - \frac{\log(N/\delta)}{\beta}(1-\gamma-\eta(1+\beta)N)$

[Choosing $\gamma, \beta, \eta \in [0,1]$ s.t. $0 \le 1-\gamma-\eta(1+\beta)N \le 1$]

$\Rightarrow$ w.p. $\ge (1-\delta)$,

$G_{i*,T} - G_T^{EXP3.P} \le \frac{\log N}{\eta} + N\beta T + \frac{\log(N/\delta)}{\beta} + \gamma G_{i*,T} + \eta(1+\beta)G_{i*,T}$

$\Rightarrow$ w.p. $\ge (1-\delta)$,

$Regret_T^{EXP3.P} \leq \frac{\log N}{\eta} + N\beta T + \frac{\log(N/\delta)}{\beta} + \gamma T + (1+\beta)\eta T$, where $G_{i^*,T} = \max_i G_{i,T} \leq T$

Now, we can optimize β, η, γ to get: $Regret_T^{EXP3.P} \leq 5.15\sqrt{TN\log(N/\delta)}$

Here optimal parameters are : $\beta = \sqrt{\frac{\log(N/\delta)}{T}}$, $\eta = 0.95\sqrt{\frac{\log N}{NT}}$, $\gamma = 1.05\sqrt{\frac{N\log N}{T}}$

$\square$

13.3 Minimax lower bound across all bandit algorithms

Here, we will work in **rewards setting** and use the same notations as in the previous section.

Theorem 13.3. *[Regret lower bound : bandits]*

inf sup $(\max_{i=1}^{N} E[\sum_{t=1}^{T} g_{i,t}] - E[\sum_{t=1}^{T} g_{I_t,t}]) \geq 1/20\sqrt{TN}$

where, infimum is over all bandit algorithms playing $(I_1, I_2, ..., I_T)$*, supremum is over all i.i.d. bernoulli reward distributions and expectation is over randomness of both rewards and algorithm.*

We will prove the theorem in the next lecture.

Bibliography

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