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

Lecture 16 — September 25

Lecturer: Aditya Gopalan *Scribe: Praveen M P*

16.1 Recap

In the last class we saw Contextual Multi Armed Bandit (C-MAB) problem which was an extension to the multi-armed bandit problem. In the C-MAB problem, apart from the MAB setting, at each time t, the learner was also allowed to observe a current 'context' $s_t \in \mathcal{S}$. We also saw that, if $|\mathcal{S}| < \infty$, looking the context at each time t, which is one of the elements of $\mathcal{S}$, one can run an EXP3 algorithm corresponding to that context.We showed that such an algorithm has upper bound on regret as

$$\text{Regret}_T^{\mathcal{S}-EXP3} \leq \sqrt{2TNlog(N)|\mathcal{S}|}$$

We also looked at a different view of C-MAB, 'Bandits with expert advice', where there are M experts and N actions. We defined the notion of pseudo-regret in this setting as

$$\text{Regret}_T := \max_{m\in[M]} \mathbb{E}\left[\sum_{t=1}^{T} l(I_t,t) - \sum_{t=1}^{T} < \pi_{t,m}, l_t >\right]$$

where $I_t \in [N]$ and $< \pi_{t,m}, l_t >$ is the average loss for each arm 'm' at t.

Treating the problem similar to a higher order bandit problem, we argued that the expected regret is of $\mathcal{O}(\sqrt{TMlog(M)})$.

In this class, we will introduce a much better algorithm, EXP4 algorithm, which is a more subtle modification of EXP3.

16.2 The EXP4 Algorithm

The EXP4 (EXP3 with expert advice) algorithm is described in Algorithm 1 EXP4. We will now show an upper bound on regret for the EXP4 algorithm.

Theorem 16.1. *EXP4 with* $\eta = \sqrt{\frac{2log(M)}{TN}}$ *gives*

$$\textit{(expected) Regret}_T \leq \sqrt{2TNlog(M)} \tag{16.1}$$

where M is the number of experts, N is the number of arms.

Note: This bound shows the power of EXP4 when N<<M and the bound is still good even when the number of experts is exponentially growing with T.

Algorithm 1 EXP4

Fix a parameter $\eta > 0$

Initialization: q_1 =Uniform$\{1, 2, ...M\}$

At each t=1,2...T

-Get expert advice $\pi_{t,1}, \pi_{t,2}, \pi_{t,3}, ...\pi_{t,M}$ each $\pi_{t,m}$ is a probability distribution over arms

-Draw $I_t \sim p_t \equiv (p_{1,t}, p_{2,t}.....p_{N,t})$ where $p_{i,t} = \sum_{m=1}^{M} q_{t,m}\pi_{t,m}(i)$

-Compute estimated losses for actions $\tilde{l}_{i,t} := \frac{l_{i,t}}{p_{i,t}}\mathbb{I}_{I_t=i}$

-Compute estimated losses for policies $\tilde{y}_{m,t} :=< \pi_{t,m}, \tilde{l}_t >$

-Update

$\tilde{Y}_{m,t} = \tilde{Y}_{m,t-1} + \tilde{y}_{m,t}$

$q_{t+1,m} := \frac{\exp(-\eta\tilde{Y}_{m,t})}{\sum_{j\in[M]} exp(-\eta\tilde{Y}_{j,t})}$

Proof: Potential function technique like in EXP3 applied to distribution q_t over expert gives

$$\sum_{t=1}^{T}\sum_{m=1}^{M} q_{t,m}\tilde{y}_{m,t} - \frac{\eta}{2}\sum_{t=1}^{T}\sum_{m=1}^{M} q_{t,m}\tilde{y}_{m,t}^2 \leq \tilde{Y}_{j,T} + \frac{Tlog(M)}{\eta} \;\; \forall j \in [M] \tag{16.2}$$

We will analyse 16.2 term by term. Consider the first term. From the definition of $\tilde{y}_{m,t}$,

$$\begin{aligned}\sum_{m=1}^{M} q_{t,m}\tilde{y}_{m,t} &= \sum_{m=1}^{M} q_{t,m} < \pi_{t,m}, \tilde{l}_t > \\ &= \sum_{m=1}^{M} q_{t,m}.\pi_{t,m}(I_t)\frac{l(I_t,t)}{\sum_{m=1}^{M} q_{t,m}.\pi_{t,m}(I_t)} = l(I_t,t)\end{aligned} \tag{16.3}$$

Along the same lines, the second term is

$$\begin{aligned}\sum_{m=1}^{M} q_{t,m}\tilde{y}_{m,t}^2 &= \sum_{m=1}^{M} q_{t,m} < \pi_{t,m}, \tilde{l}_t >^2 \\ &\leq \sum_{m=1}^{M} q_{t,m} < \pi_{t,m}, \tilde{l}_t^2 > \\ &= \sum_{m=1}^{M} q_{t,m}.\pi_{t,m}(I_t)\frac{l^2(I_t,t)}{(\sum_{m=1}^{M} q_{t,m}.\pi_{t,m}(I_t))^2} = \frac{l^2(I_t,t)}{p_{I_t,t}}\end{aligned} \tag{16.4}$$

Where the second step follows from Jensen's inequality.

Using these estimates in the potential function bound 16.2, we get

$$\sum_{t=1}^{T} l(I_t,t) - \frac{\eta}{2}\sum_{t=1}^{T}\frac{l^2(I_t,t)}{p_{I_t,t}} \leq \tilde{Y}_{j,T} + \frac{log(M)}{\eta} \tag{16.5}$$

Observe that,

$$\begin{aligned}
\mathbb{E}[\tilde{Y}_{j,T}] = \mathbb{E}_{\mathcal{F}_{t-1}}[\sum_{t=1}^{T}\mathbb{E}_{I_t}[\tilde{y}_{j,t}/\mathcal{F}_{t-1}]] &= \mathbb{E}_{\mathcal{F}_{t-1}}[\sum_{t=1}^{T}\mathbb{E}_{I_t}[<\pi_{t,j},\tilde{l}_t>/\mathcal{F}_{t-1}]] \\
&= \mathbb{E}_{\mathcal{F}_{t-1}}[\sum_{t=1}^{T}<\pi_{t,j},\mathbb{E}_{I_t}[\tilde{l}_t/\mathcal{F}_{t-1}]>] \\
&= \mathbb{E}_{\mathcal{F}_{t-1}}[\sum_{t=1}^{T}<\pi_{t,j},l_t>] \\
&= \mathbb{E}_{\mathcal{F}_{t-1}}[\sum_{t=1}^{T}y_{j,t}] = \mathbb{E}_{\mathcal{F}_{t-1}}[Y_{j,t}]
\end{aligned} \tag{16.6}$$

Where the third inequality follows from the fact that, conditioned on past, $\pi_{t,j}$ is fixed, and others follows from the definitions.
Also,

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T}\frac{l^2(I_t,t)}{p_{I_t,t}}] \leq \mathbb{E}[\sum_{t=1}^{T}\frac{1}{p_{I_t,t}}] &= \mathbb{E}_{\mathcal{F}_{t-1}}[\sum_{t=1}^{T}\mathbb{E}_{I_t}[\tfrac{1}{p_{I_t,t}}]] \\
&= \mathbb{E}_{\mathcal{F}_{t-1}}[\sum_{t=1}^{T}\sum_{i=1}^{N}p_{i,t}\tfrac{1}{p_{i,t}}] = NT
\end{aligned} \tag{16.7}$$

Taking $\mathbb{E}[.]$ on both sides of 16.5 , using 16.6 , 16.7 rearranging terms, we get expected regret as

$$\text{Regret}_T^{\text{EXP4}} \leq \frac{log(M)}{\eta} + \frac{\eta}{2}TN \tag{16.8}$$

It is easy to show that the optimum value of η that minimises the RHS is $\eta = \sqrt{\frac{2log(M)}{TN}}$ and hence the regret bound is

$$\text{(expected) Regret}_T \leq \sqrt{2TNlog(M)} \tag{16.9}$$

□

16.3 Stochastic Multi-Armed Bandits

So far, we have discussed about the setting with adversarial losses (outcomes) of arms (actions).

The main criticisms of the discussed model can be
(i) No structure was assumed on losses (Real life systems might have probabilistic structure)

(ii) Worst case guarantees derived might be too pessimistic (It might be okay to do well in average case over losses/outcomes)
Consider the setup where:
There is a set of random variables $\{X_{i,s} : i = 1, 2, ...N; s = 1, 2, ...\}$
Where, $X_{i,s}$ is the reward from arm i, when played s^{th} successive time.

We assume:
• All the $\{X_{i,s}\}_{i,s}$ are independent
•$\forall$ i, $\{X_{i,1}, X_{i,2}, X_{i,3}....\}$ are identically distributed with $\mathbb{E}[X_{i,s}] = \mu_i$; and $\mu^* = \max_{i\in[N]} \mu_i$

At each time t, we want to play an arm $I_t \in$ [N], based on (I_1,I_2,....,I_{t-1})and associated rewards.
Goal: The goal is to maximize Total (expected) Reward in T plays. ie,

$$\text{Maximize: } \mathbb{E}[\sum_{t=1}^{T} \mu_{I_t}] = \sum_{i=1}^{N} \mu_i \mathbb{E}[T_i(T)] \quad \text{where } T_i(t) := \sum_{s=1}^{T} \mathbb{I}_{I_s=i}$$

$$\text{Equivalently, } \quad \text{Minimize: } \{\mu^* T - \sum_{i=1}^{N} \mu_i \mathbb{E}[T_i(T)]\} \tag{16.10}$$

(where μ^* is the mean corresponding to the best arm in the average sense.)
We call 16.10 as psuedo regret.

Note:

$$\begin{aligned} \bullet \max_{i\in[N]} \mathbb{E}[\mu_i(T)] - \mathbb{E}[\sum_{j=1}^{N} \mu_j T_j(T)] &\leq \mathbb{E}[\max_{i\in[N]} \mu_i(T)] - \mathbb{E}[\sum_{j=1}^{N} \mu_j T_j(T)] \\ &= \mathbb{E}[\max_{i\in[N]} \mu_i(T) - \sum_{j=1}^{N} \mu_j T_j(T)] \end{aligned} \tag{16.11}$$

•Can run EXP3 to get (pseudo) Regret=$\mathcal{O}(\sqrt{NTlog(N)})$. But EXP3 doesn't exploit the structure of rewards.

16.3.1 Naive Approach

Since the rewards are stochastic in nature and empirical estimates are unbiased estimates of the true mean, a naive approach can be the 'greedy' approach (Follow The Leader) as described in Algorithm 2 (FTL). As an example consider the setting of two-armed bandits, each having reward Bernoulli(μ_i), i=1,2 and $1 > \mu_1 > \mu_2 > 0$. So, the arm 1 is better in expected sense, and thus is the arm that we want to learn. Consider the algorithm:

It is easy to see that this algorithm is bad, and gives regret $\mathcal{O}$(T). This is because, say if both of the arms were played once to estimate the empirical mean, the probability that the event $\{X_{1,1}=0, X_{2,1}=1\}$ occurs is (1-μ_1)(μ_2) > 0. Hence Regret$_T \geq (1-\mu_1)(\mu_2)T(\mu_1 - \mu_2)$

Algorithm 2 FTL

Initialization: Play each arm for a fixed number of times.

$$\text{Select the best arm } \hat{i} = \underset{i\in\{1,2\}}{\operatorname{argmax}} \left[\frac{\sum_{s=1}^{T_i(t)} X_{i,s}}{T_i(t)} \right]$$

Subsequently:Always play arm $\hat{i}$

16.3.2 The ε-FIRST Algorithm

The ε-FIRST Algorithm solves the 'linear regret problem' by adding an 'exploration' phase, before going to greedy phase.

Algorithm 3 ε-FIRST

Fix a parameter $0 \leq \varepsilon \leq 1$

Phase 1: Play all arms round robin upto time εT.

Phase 2: For t$\geq \varepsilon$T+1 play arm $\mathrm{I}_t = \underset{i}{\operatorname{argmax}} \ \hat{\mu}_i(\varepsilon T)$

Theorem 16.2. *The regret bound for the $\varepsilon-$ FIRST algorithm for the optimum value of ε for a time horizon of T is*

$$\textit{Regret}_T^{\varepsilon FIRST} \leq 1 + \frac{2\,N}{\Delta^2} log(2NT) \tag{16.12}$$

where $\Delta = \min_i (\mu^* - \mu_i)$

Proof: From the definition 16.10,

$$\begin{aligned}
\text{Regret}_T &= \mu^*\mathrm{T} - \sum_{i=1}^{N} \mu_i \mathbb{E}[\mathrm{T}_i(\mathrm{T})] \\
&= \mu^*\varepsilon\mathrm{T} - \sum_{i=1}^{N} \mu_i \mathbb{E}[\mathrm{T}_i(\varepsilon\mathrm{T})] + \mu^*(1-\varepsilon)\mathrm{T} - \sum_{i=1}^{N} \mu_i \mathbb{E}[\mathrm{T}_i(\mathrm{T}) - \mathrm{T}_i(\varepsilon\mathrm{T})] \\
&= \text{Regret [Phase 1] + Regret [Phase 2]}
\end{aligned} \tag{16.13}$$

Now, since $\mathrm{X}_{i,s} \in [0,1]$, Regret [Phase 1] $\leq \varepsilon T$

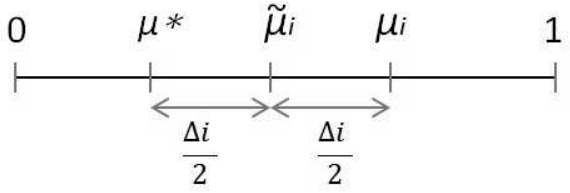

Figure 16.1. An illustration of the mean rewards.

$$\begin{aligned}
\text{Regret}[\,\text{Phase2}\,] &= \mu^*(1-\varepsilon)\text{T} - \sum_{i=1}^{N} \mu_i(1-\varepsilon)\,\text{T}\,\mathbb{E}[\text{ no: of times arm i was played in phase 2}] \\
&= \mu^*(1-\varepsilon)\text{T} - \sum_{i=1}^{N} \mu_i(1-\varepsilon)\,\text{T}\,\mathbb{P}r\left[\text{I}_{\varepsilon T+1} = i\right] \\
&= (1-\varepsilon)\text{T} \sum_{i=1,\, i\neq i^*}^{N} (\mu^* - \mu_i)\,\mathbb{P}r\left[\text{I}_{\varepsilon T+1} = i\right] \;\leq\; (1-\varepsilon)\text{T}\,\mathbb{P}r\left[\text{I}_{\varepsilon T+1} \neq i\right] \quad (16.14)
\end{aligned}$$

Now, fix some i $\neq i^*$. Let $\Delta_i = \mu^* - \mu_i$ and $\Delta = \min_i \Delta_i$ as ilustrated in Figure 16.1

$$\begin{aligned}
\mathbb{P}r\left[\text{I}_{\varepsilon T+1} = i\right] &\leq \mathbb{P}r\left[\hat{\mu}_{i(\varepsilon T)} > \hat{\mu}_{i^*(\varepsilon T)}\right] \\
&\leq \mathbb{P}r\left[\hat{\mu}_{i(\varepsilon T)} > \mu_i + \frac{\Delta_i}{2}\right] + \mathbb{P}r\left[\hat{\mu}_{i(\varepsilon T)} < \mu^* - \frac{\Delta}{2}\right] \\
&\leq \mathbb{P}r\left[\hat{\mu}_{i(\varepsilon T)} > \mu_i + \frac{\Delta}{2}\right] + \mathbb{P}r\left[\hat{\mu}_{i(\varepsilon T)} < \mu^* - \frac{\Delta}{2}\right] \\
&\leq 2\exp\left(\frac{-2\varepsilon T}{N}\left(\frac{\Delta}{2}\right)^2\right) \quad (16.15)
\end{aligned}$$

$$\text{Also } \mathbb{P}r\left[\text{I}_{\varepsilon T+1} \neq i\right] \leq 2\,N\exp\left(\frac{-\varepsilon T\Delta^2}{2\,N}\right) \quad (16.16)$$

Where the inequalities 16.15 and 16.16 follows from Azuma-Hoeffding Inequality (Theorem 4.2). Therefore, putting 16.16 in 16.14 and using 16.13

$$\text{Regret}_T^{\varepsilon Fixed} \leq \varepsilon\, T + 2N\exp\left(\frac{-\varepsilon\Delta^2}{2\,N}\right) \quad (16.17)$$

Set the optimum value of $\varepsilon = \frac{2}{\Delta^2}\frac{N}{T}log(2\,N\,T)$ to get

$$\text{Regret}_T^{\varepsilon FIRST} \leq 1 + \frac{2\,N}{\Delta^2}log(2NT) \quad (16.18)$$

□

Note: Δ has to be known in prior to fix the value of the parameter ε

16.4 References

[1] Chapter-4, Sébastien Bubeck and Nicolò Cesa-Bianchi, " Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems."