E1 245: Online Prediction & Learning	Fall 2014

Lecture 17 — September 30

Lecturer: Aditya Gopalan *Scribe: Raj Kumar Maity*

17.1 RECAP

In previous class we have seen that for a **Stochastic MAB**
Setup: N arms of a bandit. Random variables $\{X_{i,s} : i = [N], s = 1, 2\ldots$
$X_{i,s}$ = reward for arm i when played for the s-th successive time . All $\{X_{i,s}\}_{i,s}$ are independent
$\forall i, \{X_{i,1}, X_{i,2}\ldots\}$ identically distributed. $\mathbb{E}[X_{i,s}] = \mu_i, \mu* = \max_{i=1}^{N} \mu_i$
We can run EXP-3 algo to get Regret $= O(\sqrt{NT \log N})$ but doesn't exploit the stochastic structure of rewards.Playing the FTL(greedy) gives linear regret(bad).Adding a exploration phase before "GREEDY"(ε first algorithm with $0 \le \varepsilon \le 1$) we get

$$Regret_T \le 1 + \frac{2N}{\Delta^2}\log(2NT)$$
$$setting\ \varepsilon = \frac{2N}{\Delta^2 T}\log(2NT)$$
$$where\ \Delta = \mu^* - \mu_i; \Delta = \min_i \Delta_i \qquad Drawback: Need\ to\ know\ \Delta.$$

In this class we will see an elegant algorithm to fix the problem.

17.2 Stochastic Multiarm Bandit

Algorithm(UCB-"upper confidence bound ")

- initially play each arm $i \in [N]$ once

for $t \ge N+1$

{

1. play arm $arg\max_{i\in[N]}(\hat{\mu}_i(t) + C_{t,T_i(t)})$,where
$\hat{\mu}_i(t) = \frac{1}{T_i(t)}\sum_{s=1}^{T_i(t)} X_{i,s}$ (empirical mean)
$T_i(t) = \sum_{s=1}^{t} \mathbb{I}(I_s = i)$
$C_{t,s} = \sqrt{\frac{2\log t}{s}}\ for\ t, s \ge 0;$

}

Theorem 17.1. *[1] Suppose* $X_{i,s} \in [0,1]$ *then* $Regret_T^{UCB} \leq 8\log T(\sum_{i\neq i^*}\frac{1}{\Delta_i}) + \frac{\pi^2}{3}\sum_{i=1}^{N}\Delta_i$
$\{\Delta_i = \mu^* - \mu_i, \mu^* = max_i\mu_i\}$

Proof: Suppose we are on round $t \leq T$.For any suboptimal arm $i \neq i^*$

$$\begin{aligned}\mathbb{P}[\hat{\mu}_i(t) - C_{t,T_i(t)} \geq \mu_i] &= \mathbb{P}[\hat{\mu}_i(t) - \mu_i \geq C_{t,T_i(t)}] \\ &\leq \sum_{n=1}^{t}\mathbb{P}[\frac{1}{n}\sum_{s=1}^{n}X_{i,s} - \mu_i \geq C_{t,n}] \\ &\leq \sum_{n=1}^{t} exp(-2n\frac{2\log t}{n}) = t^{-3}\end{aligned}$$

$\{$*Hoeffding inequality :*$\{Y_i\}$*iid random variable* $Y_i \in [0,1]$ $\quad \mathbb{P}[\frac{1}{n}\sum_{i=1}^{n}Y_i - \mathbb{E}[Y_i] > \varepsilon] \leq exp(-2n\varepsilon^2)\}$

Similarly $\mathbb{P}[\hat{\mu_{i^*}}(t) + C_{t,T_{i^*}(t)} \leq \mu^*] \leq t^{-3}$

With probability $\geq 1 - 2t^{-3}$,

$$\{\hat{\mu}_i(t) - C_{t,T_{i(t)}} < \mu_i = \mu^* - \Delta_i < \hat{\mu}_{i^*}(t) + C_{t,T_{i^*}(t)} - \Delta_i\} \qquad -\{\textit{call this event } A\}$$

$$\{\textit{If } T_i(t) \geq \frac{8\log T}{\Delta_i^2}(\geq \frac{8\log t}{\Delta_i^2}) \Rightarrow \Delta_i - 2\sqrt{\frac{2\log t}{T_i(t)}} \geq 0\}$$

$$\begin{aligned}&\Longleftrightarrow [\hat{\mu}_i + C_{t,T_{i(t)}}] + [\Delta_i - 2C_{t,T_{i(t)}}] < \hat{\mu}_{i^*}(t) + C_{t,T_{i^*}(t)} \\ &\Rightarrow \hat{\mu}_i + C_{t,T_{i(t)}} < \hat{\mu}_{i^*}(t) + C_{t,T_{i^*}(t)} \Rightarrow I_t \neq i\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}[T_i(t)] =& \mathbb{E}[1+\sum_{t=N+1}^{T}\mathbb{I}\{I_t=i\}] \\
&\leq \mathbb{E}[l_i+\sum_{t=1}^{T}\mathbb{I}\{I_t=i,T_i(t)\geq l_i\}] && \{l_i=\frac{8\log T}{\Delta_i^2}\} \\
&= l_i+\sum_{t=1}^{T}]\mathbb{P}\{I_t=i,T_i(t)\geq l_i\}] && \{A\cap\{T_i(t)=l_i\}\subseteq\{I_t\neq i\}\} \\
&\leq l_i+\sum_{t=1}^{T}\mathbb{P}[A_t^c] \\
&\leq \frac{8\log T}{\Delta_i^2}+\sum_{t=1}^{\infty}\frac{2}{t^3} \\
&\leq \frac{8\log T}{\Delta_i^2}+2\sum_{t=1}^{\infty}\frac{1}{t^2} \\
&= \frac{8\log T}{\Delta_i^2}+\frac{\pi^2}{3}
\end{aligned}
$$

□

NOTE: Optimal regret scaling for stochastic bandits is (Lai and Robbins 1985 [4])

$$\liminf_{T\to\infty}\frac{\mathbb{E}[T_i(T)]}{\log T}\geq\frac{1}{D(\mu_i||\mu^*)}$$

-achieved by KL-UCB[2].

17.2.1 Pure Exploration in Stochastic Bandit

Motivation:

1. Suppose there is a budget of plays for experimentation.
2. Regret penalizes *every* suboptimal play, but this may not be desirable when there is an experimentation budget.

GOAL: Identify the best arm in a Bandit as quickly as possible.
-Sequential hypothesis testing but with the flexibility of picking 1 arm each time.
Defination :A bandit algorithm {i.e a rule mapping history of plays to arms } is called an (ε,δ) PAC algorithm (Probably approximately corectly) with sample complexity T if

1. It outputs a ε optimal arm with probability $\geq(1-\delta)$ when it terminates.
2. No of time steps taken to terminate $\leq T$

NOTE:

1. Fixed confidence setting i.e fix δ.
2. Fixed budget setting i.e fix no of plays.

Naive Algorithm (-uniformly sample all arms.)

-Parameter (ε, δ)
-N arms ; i $\in \{1, 2 \dots N\}$

1. Sample each arm i for $l = \frac{2}{\varepsilon^2}\log(\frac{2N}{\delta})$ times
2. Let $\hat{\mu}_i$ be its emperical mean.
3. output $i' = argmax_i \hat{\mu}_i$

Theorem 17.2. *Naive (ε, δ) is (ε, δ) -PAC algorithm with sample complexity $\frac{2N}{\varepsilon^2}\log\frac{2N}{\delta}$.*

Proof: We will show that its (ε, δ) -PAC algorithm.
Let i be an arm such that $\mu_i < \mu^* - \varepsilon$

$$\begin{aligned}
\mathbb{P}[\hat{\mu}_i > \hat{\mu}_{i^*}] &\leq \mathbb{P}[\{\hat{\mu}_i > \mu_i + \frac{\varepsilon}{2}\} \cup \{\mu_{i^*} < \mu^* - \frac{\varepsilon}{2}\}] \\
&\leq 2exp[-2l(\frac{\varepsilon}{2})^2] \\
&= \frac{\delta}{N}
\end{aligned}$$

Summing over all such i,
$\mathbb{P}$[alg fails to output ε optimal arm]$\leq \frac{\delta}{N}N = \delta$ □

Improvement:$O(\frac{N}{\varepsilon^2}\log\frac{N}{\delta}) \to O(\frac{N}{\varepsilon^2}\log\frac{1}{\delta})$

17.2.2 Median Elimination

Idea:Eliminate bad arm in phases
Algorithm:

-Parameter (ε, δ)
-initialize :$S_1 = \{1, 2, \dots N\}, \varepsilon_1 = \frac{\varepsilon}{4}, \delta_1 = \frac{\delta}{2}, l = 1$
Untill($|S_l| = 1$):
{

1. sample each arm in S_l for $n_l = \frac{1}{(\frac{\varepsilon_l}{2})^2}\log\frac{3}{\delta_l}$ times .
   Let $\hat{\mu}_{i,l}$ denote the resulting emperical mean .
2. Let $m_l = MEDIAN(\hat{\mu}_{i,l} : i \in S_l)$;
   $S_{l+1} = S_l \setminus (i : \hat{\mu}_{i,l} < m_l)$
3. $\varepsilon_{l+1} = \varepsilon_l; \delta = \delta_l; l = l+1.$

}

Theorem 17.3. *[3] Median elimination is (ε,δ) -PAC algorithm with sample complexity $O(\frac{N}{\varepsilon^2}\log\frac{1}{\delta})$*

let's show that at the l-th phase the expected reward of the best surviving arm from S_l drops by at most ε_l with probability $\geq (1-\delta_l)$
LEMMA 1: for every phase l , $\mathbb{P}[\max_{j\in S_l}\mu_j \leq \max_{i\in S_{l+1}}\mu_i + \varepsilon_l] \geq (1-\delta_l)$

Proof: Without loss of generality , lets consider l=1 , $max_{i\in S_l}\mu_i = \mu^* = \mu_{i^*}$

$$Let\ E_1 = \{\hat{\mu_{i^*}} < \mu^* - \frac{\varepsilon}{2}\}$$
$$\mathbb{P}[E_1] \leq exp(-2n_l(\frac{\varepsilon}{2})^2) \leq \frac{\delta_1}{3}$$

Now lets take an arm j that is not ε_1 optimal

$$\begin{aligned}
&\mathbb{P}[\hat{\mu}_j > \hat{\mu_{i^*}} | E_1^c] \\
&\leq \mathbb{P}[\hat{\mu}_j > \hat{\mu_{i^*}} - \frac{\varepsilon}{2} | E_1^c] \\
&\leq \mathbb{P}[\hat{\mu}_j > \mu_j + \frac{\varepsilon}{2} | E_1^c] \\
&\leq \mathbb{P}[\hat{\mu}_j > \mu_j + \frac{\varepsilon}{2}] \\
&\leq \frac{\delta_1}{3} && \{Hoeffding\ Inequality\}
\end{aligned}$$

Let B be the no of arms j which are not ε_1 optimal but emperically better than i^*

$$\mathbb{E}[B|E_1^c] \leq \frac{N\delta_1}{3}$$
$$\textit{By Markov inequality: } \mathbb{P}[B \geq \frac{N}{2}|E_1^c] \leq \frac{\frac{N\delta_1}{3}}{\frac{N}{2}} = \frac{2\delta_1}{3}$$
$$\mathbb{P}[B \geq \frac{N}{2}] \leq \mathbb{P}[B \geq \frac{N}{2}|E_1^c] + \mathbb{P}[E_1]$$
$$\leq \frac{2\delta_1}{3} + \frac{\delta_1}{3} = \delta_1$$
$$\mathbb{P}[B < \frac{N}{2}] \geq 1 - \delta_1.$$

□

<u>LEMMA 2</u>:Sample complexity $O(\frac{N}{\varepsilon^2}\log\frac{1}{\delta})$

Proof: In phase l , total no of samples $= \frac{4}{\varepsilon_l^2}|S_l|\log(\frac{3}{\delta_l})$
Sample complexity

$$= \sum_{l=1}^{\log N} \frac{4}{\varepsilon_l^2}|S_l|\log(\frac{3}{\delta_l}) = O(\frac{N}{\varepsilon^2}\log\frac{1}{\delta})$$
$$\{|S_l| = \frac{N}{2^l}; \varepsilon_l = \frac{\varepsilon}{4}(\frac{3}{4})^{l-1}; \delta_l = \frac{\delta}{2^l}\}$$

with probabilty $\geq (1-\delta_l)$,best mean $\geq$ previous best mean $-\varepsilon_l$
By union bound,
with prob $\geq 1 - \underbrace{(\delta_1 + \delta_2 + \ldots + \delta_{\log N})}_{\leq \delta}$,
best mean $\geq \mu^* - \underbrace{(\varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_{\log N})}_{\leq \varepsilon}$

□

Bibliography

[1] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[2] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, Gilles Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[3] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7:1079–1105, 2006.

[4] TL Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.