ECE-GY 9223 Reinforcement Learning Spring 2019

Lecture 9 — May 3, 2019

Prof. Quanyan Zhu *Scribe: Arihant Jain*

1 Overview

In the last lecture we discussed TD learning, average cost problems, resulting ACOE and the shifted our attention stochastic and adversarial bandit problems.

In this chapter in order to continue our discussion on these bandit problems we stat by studying Hoeffding bound and and Hoeffding lemma. Both these tools will be useful to do regret analysis on these bandits. This shall be he order of our discussion:

- Hoeffding Bounds
- Hoeffding Lemma
- Regret analysis for stochastic bandits
- Regret analysis for adversarial bandits

2 Hoeffding Bound

Theorem 1. *Given a sequence of independdant random variables $\{X_i\}$, where X_i takes values between $[a_i, b_i]$*

$$P\left(\left|\frac{\sum_i X_i - \mathbb{E}(\sum_i X_i)}{n}\right| \geq x\right) \leq 2\exp\left(-\frac{2n^2x^2}{\sum_i (b_i - a_i)^2}\right)$$

Proof. For iid bernauli random variable
let X_i be iid Bernoulli random variables with parameter p

$$\begin{aligned} Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq p + x\right) &\leq e^{n\theta(p+x)}\,\mathbb{E}(e^{\theta\sum_{i=1} nX_i}) \\ &= \mathbb{E}(\theta X_i)^n e^{-n\theta(p+x)} \\ &= (e^{\theta} + (1-p))^n e^{-n\theta(p+x)} \\ &= e^{n\log(pe^{\theta}+1-p)} e^{-n\theta(p+x)} \end{aligned} \tag{1}$$

$$f(\theta) = \log(pe^{\theta} + 1 - p) \equiv f(0) + f'(0)\theta + \frac{1}{2}f''(u)\theta^2 \text{ for } u \in [0, \theta]$$

$$\begin{aligned}
f(0) &= 0 \\
f'(0) &= \frac{pe^\theta}{pe^\theta + 1 - p}\bigg|_{\theta=0} = p \\
f''(\theta) &= \frac{(pe^\theta + 1 - p)pe^\theta - (pe^\theta)^2}{(pe^\theta + 1 - p)^2} = \left(\frac{pe^\theta}{pe^\theta + 1 - p}\right)\left(\frac{1-p}{pe^{theta} + 1 - p}\right) \\
&\leq \frac{1}{4} \\
\Rightarrow f(\theta) &\leq p\theta + \frac{1}{8}\theta^2 \qquad (2)
\end{aligned}$$

Using equation 1 and 2

$$\begin{aligned}
e^{n\log(pe^\theta+1-p)}e^{-n\theta(p+x)} &\leq e^{-n(p+x)\theta}e^{n(p\theta+\frac{\theta^2}{8})} \\
&= e^{-nx\theta+\frac{1}{8}n\theta^2} \qquad (3) \\
&\leq e^{2nx^2} \qquad (4) \\
Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \leq p - x\right) &\leq e^{-2nx^2}
\end{aligned}$$

To go from equation 3 to 4 we use calculate the $inf - nx\theta + \frac{1}{8}n\theta^2$ for $\theta > 0$; which occurs at $\theta = 4x$, which can proved trivially. □

2.1 Accuracy vs Confidence

We will use the bound we just proved to look at this question. "How many samples do we need to get sufficiently close to the mean ?"
Let $0 \leq Z_i \leq 1$ be iid distributed random variables. Then from our theorem it follows that:

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}(Z_i)\right| > \epsilon\right) \leq 2\exp(-2n\epsilon^2)$$

ϵ can be thought of as accuracy and $2\exp(-2n\epsilon^2)$ as confidence
Accuracy means how far we can allow the sample mean to be from the true mean, confidence means with what probability can we allow this to happen

Lemma 2. *Hoeffding Lemma, suppose we choose*

$$n \geq \frac{1}{2\epsilon^2}\log\frac{2}{\delta}$$

then with probability at least $1 - \delta$, the difference between the empirical mean $\frac{1}{n}\sum_{i=1}^n Z_i$ and the true mean $\mathbb{E}(Z_i)$ is at most ϵ.

Proof. Suppose

$$n \geq \frac{1}{2\epsilon^2}\log 2\delta$$

Then, by Hoeffding's inequality,

$$\begin{aligned}
Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} - \mathbb{E}(Z_i)\right| > \epsilon\right) &\leq 2\exp(-2n\epsilon^2) \\
&\leq 2\exp(-2\frac{1}{2\epsilon^2}\log(\frac{2}{\delta})\epsilon^2) \\
&= 2\exp\left(-\log\frac{2}{\delta}\right) \\
&= \delta
\end{aligned}$$

□

Note that, accuracy is expensive, while confidence is cheap. SUppose we find n for some ϵ,δ, and then we decide that we want 10 times more confidence. We can calculate that for $\delta' = \frac{\delta}{10}$, we will need

$$n' = \frac{1}{2}\left(\frac{1}{\epsilon}\right)^2 \log\frac{2\cdot 10}{\delta} = n + \frac{1}{2}\left(\frac{1}{\epsilon}\right)^2 \log(10) = n + C(\epsilon) \tag{5}$$

samples to achieve this. We can just add a constant number $C(\epsilon)$of extra samples. If we would like 100 times more confidence, we can just add $2C(\epsilon)$ extra samples. Or we can just say $n \propto \log\frac{2}{\delta}$ or $\delta \propto \frac{2}{\exp(n)}$.

On the other hand, accuracy is quite expensive. Suppose that we decide we want 10 times more accuracy. We can calculate that for $\epsilon' = \frac{\epsilon}{10}$, we will need $100n$ samples. An increase of a factor of 100! Or we can just say $\epsilon \propto \frac{1}{\sqrt{n}}$.

Lemma 3. *Hoeffding Lemma: let X be any real valued random variable with $\mathbb{E}(X) = 0$. $a \leq X \leq b$ almost surely. Then for all $\lambda \in R$.*

$$\mathbb{E}(e^{\lambda X} \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right))$$

3 Regret analysis for stochastic multi arm bandit

For a stochastic multi arm bandit with $K \geq 2$ arms, the agent chooses an arm I_t at each time step $t = 1, 2, 3, ...$ from the set of arms $\{1, 2, ..., K\}$ and obtains reward drawn from V_{I_t} independently from the past outcome. In the last lecture we defined regret as

$$\begin{aligned}
R_n &= n\mu^\star - \sum_{t=1}^{n} X_t \\
\mathbb{E}(R_n) &= n\mu^\star - \mathbb{E}[\sum_{t=1}^{n} \mu_{I_t}]
\end{aligned} \tag{6}$$

Here $\mu_k = \int X dV_k(X)$ is the mean reward of arm k, and $\mu^\star = \max_{k \in 1,2,\ldots,K} \mu_k$
Now we define $T_k(n) = \sum_{t=1}^{n} \mathbb{1}\{I_t = k\}$ i.e. total number of times arm k was pulled from round one to round n and $\sum_{k=1}^{K} T_k(n) = n$; which gives us

$$\mathbb{E}(R_n) = \mathbb{E}[\sum_{k=1}^{K} T_k(n)(\mu^\star - \mu_k)]$$

$(\mu^\star - \mu_k)$ is Δ_k (gap) which measures the sub optimality of an arm

$$I_t = \arg\max_k \frac{1}{T_k(t-1)} \sum_{i=1}^{T_k(t-1)} X_{k,i} + \sqrt{\frac{2\log t}{T_k(t-1)}}$$

Analysis

$$P\left(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \geq \epsilon\right) \leq \exp(-2s\epsilon^2)$$
$$P\left(\frac{1}{s}\sum_{i=1}^{s} X_i - \mu \leq -\epsilon\right) \leq \exp(-2s\epsilon^2)$$

we pick $\epsilon = \sqrt{\frac{2\log t}{s}}$ and setting $\frac{1}{s}\sum_{i=1}^{s} X_i$ to $\hat{\mu}_{k,s}$, $\hat{\mu}_{k,s}$ is the average reward we get from pulling arm k in s rounds; the corresponding equations become after rearranging

$$P\left(\hat{\mu}_{k,s} - \sqrt{\frac{2\log t}{s}} \geq \mu_k\right) \leq t^{-4}$$
$$P\left(\hat{\mu}_{k,s} + \sqrt{\frac{2\log t}{s}} \leq \mu_k\right) \leq t^{-4}$$

Proposition 4. *Each sub-optimal arm k is pulled on average at most*

$$\mathbb{E}(T_k(n)) \leq \frac{8\log n}{\Delta_k^2} + \frac{\pi^2}{3}$$

times. As a result

$$\begin{aligned}\mathbb{E}(R_n) &= \sum_k \Delta_k \, \mathbb{E}(T_\mu(n)) \\ &\leq 8 \sum_{k,\Delta_k \geq 0} \frac{\log n}{\Delta_\mu} + \frac{k\pi^2}{3}\end{aligned}$$

which gives us; $\forall s, t, k$

$$\mu_k - \sqrt{\frac{2\log t}{s}} \leq \hat{\mu}_{k,s} \qquad w.p \quad 1 - t^{-4}$$
$$\mu_k + \sqrt{\frac{2\log t}{s}} \geq \hat{\mu}_{k,s} \qquad w.p \quad 1 - t^{-4}$$

If a sub-optimal arm is pulled at time t

$$\hat{\mu}_{k,T_k(t-1)} + \sqrt{\frac{2\log t}{T_\mu(t-1)}} \geq \hat{\mu}_{k^\star,T_{k^\star}(t-1)} + \sqrt{\frac{2\log t}{T_{\mu^\star}(t-1)}}$$

we have

$$\mu_k + 2\sqrt{\frac{2\log t}{T_\mu(t-1)}} \geq \hat{\mu}_{k,T_k(t-1)} + \sqrt{\frac{2\log t}{T_\mu(t-1)}} \geq \hat{\mu}_{k^\star,T_{k^\star}(t-1)} + \sqrt{\frac{2\log t}{T_{\mu^\star}(t-1)}} \geq \mu_{k^\star}$$

and

$$T_\mu(t-1) \geq \frac{8\log t}{(\mu^\star - \mu_k)^2} = \frac{8\log t}{\Delta_k^2}$$

4 Regret Analysis for Adversarial Bandit problem

The characteristic of an adversarial bandit is that the rewards are chosen by arbitrarily by an adversary at $t = 1, 2, 3, ..., n$; the adversary selects $X_t(1), ..., X_t(k) \in$ [0,1] and the player selects an arm I_t (according to some strategy adapted to the information)
We can imagine two scenarios:

- Full information: The player observes the rewards for all the arm $X_t(k)$ for $k \in 1, 2, 3, ..., K$

- Bandit Information: The player observes the reward of the selected arm

We define the regret as

$$R_n(k) = \sum_{t=1}^{n} X_t(k) - \sum_{t=1}^{n} X_t(I_t)$$

Goal : find an algorithm that generates low regret for all reward sequences

4.1 Exponentially Weighted Forecaster

We will consider the case where we have full information, which gives us Exponentially Weighted Forecaster
For this case we have

- $w_1(k) = 1$ for all arms

- $t = 1, ..$ the player selects an arm$I_t \sim P_t$

$$P_t(k) = \frac{w_t(k)}{\sum_{i=1}^{k} w_t(i)}$$

where $w_t(k) = e^{\eta \sum_{s=1}^{t-1} X_s(k)}$ for $\eta > 0$

Proposition 5. *Let $\eta \leq 1$. Then the regret of EWF is bounded as*

$$\begin{aligned} R_n &\leq \frac{\log k}{\eta} + \frac{\eta n}{8} \\ R_n &\leq \sqrt{\frac{n \log k}{2}} \qquad \textit{if} \quad \eta = \sqrt{\frac{8 \log k}{n}} \end{aligned}$$

Proof. Let $W_t = \sum_{k=1}^{K} w_t(k)$ sum over all arms

$$\begin{aligned} \frac{W_{t+1}}{W_t} &= \frac{\sum_{k=1}^{K} w_{t+1}(k)}{W_t} = \frac{e^{\eta \sum_{s=1}^{t} X_s(k)}}{W_t} \\ &= \sum_{k=1}^{K} \frac{W_t(k) e^{\eta X_t(k)}}{W_t} \\ &= \sum k = 1^K P_t(k) e^{\eta X_t k} \\ &= \mathop{\mathbb{E}}_{I \sim P_t} [e^{\eta X_t(I)}] \\ &= \mathop{\mathbb{E}}_{I \sim P_t} [e^{\eta (X_t(I) - \mathbb{E}_{J \sim P_t}(X_t(j)))}] e^{\eta \, \mathbb{E}(X_t(J))} \\ &\leq e^{\frac{\eta^2}{8}} e^{\eta \, \mathbb{E}(X_t(I))} \qquad \text{by Hoeffding bound} \end{aligned}$$

where $\mathbb{E}_{J \sim P_t}[X_t(J)]$ is the mean of $X_t(J)$. Hence

$$\begin{aligned} \log \frac{W_{t+1}}{W_t} &\leq \eta \, \mathbb{E}(X_t(J)) + \frac{\eta^2}{8} \\ \log \frac{W_{n+1}}{W_1} &\leq \eta \, \mathbb{E}[\sum_{t=1}^{n} X_t(J)] + \frac{n \eta^2}{8} \end{aligned}$$

and

$$\begin{aligned} \log \frac{W_{t+1}}{W_t} &= \log \frac{\sum_{k=1}^{K} e^{\eta \sum_{t=1}^{n} X_t(k)}}{\sum_{k=1} K W_1(k)} \\ &= \log \sum_{k=1}^{K} e^{\eta \sum_{t=1}^{n} X_t(k)} - \log K \\ &\geq \eta \sum_{t=1}^{n} X_t(k) - \log k \qquad \forall k = 1, 2, .., K \end{aligned}$$

using the above 2 results we get

$$\forall k \quad \mathbb{E}[R_n(k)] = \sum_{t=1}^{n} X_t(h) - \mathbb{E}[\sum_{t=1}^{n} X_t(I_t)] \leq \frac{\log K}{\eta} + \frac{n\eta}{8}$$

□

4.2 Exploration Exploitation using Exponential Weights (EXP3)

For the case where we just have bandit information we come up with EXP3. This case is characterized by

- $W_1(h) = 1$ for all $k = 1, 2, ..., K$
- pick $I_t \sim P_t$

where

$$P_t(h) = (1-\beta)\frac{W_t(k)}{\sum_{i=1}^{k} W_t(i)} + \frac{\beta}{K}$$

here $W_t(k) = e^{\eta \sum_{s=1}^{t-1} \hat{X}_s(k)}$, $\hat{X}_s(k) = \frac{X_s(k)}{P_s(k)} \mathbb{1}\{I_s = k\}$ and $\eta > 0, \beta > 0$ are the parameters of the algorithms

Proposition 6. *Let $\eta \leq 1$ and $\beta = \eta K$. Then the regret of EXP3 is bounded as*

$$R_n \leq \frac{\log K}{\eta} + (e-1)\eta n K$$

If we select $\eta = \sqrt{\frac{\log K}{(e-1)nK}}$, the regret of EXP3 is bounded as

$$R_n \leq 2.63\sqrt{nK \log K}$$

Proof. Let $W_t = \sum_{k=1}^{K}$. Note that $\mathbb{E}_{I_s \sim p_s}[\hat{X}_s(k)] = \sum_{i=1}^{K} p_s(i) \frac{X_s(k)}{p_s(k)} \mathbb{1}\{i = k\} = X_s(k)$ and $\mathbb{E}_{\hat{X}_s(I_s)} = \sum_{i=1}^{K} p_s(i) \frac{X_s(i)}{p_s(i)} \leq K$. Therefore, we have

$$\begin{aligned}
\frac{W_{t+1}}{W_t} &= \sum_{k=1}^{K} \frac{w_t(k) e^{\mu \hat{X}_i(k)}}{W_t} = \sum_{k=1}^{K} \frac{p_t(k) - \beta/K}{1-\beta} e^{\mu \hat{X}(k)} \\
&\leq \sum_{k=1}^{K} \frac{p_t(k) - \beta/K}{1-\beta} (1 + \eta \hat{X}_t(k) + (e-2)\eta^2 \hat{X}_t(k)^2) \\
&\leq 1 + \frac{1}{1-\beta} \sum_{k=1}^{K} p_t(k)(\eta \hat{X}_t(k) + (e-2)\eta^2 \hat{X}_t(k)^2)
\end{aligned}$$

This is because $\eta \hat{X}_t(k) \leq \eta K/\beta$ and $e^x \leq 1 + x + (e-2)x^2$ for $x \leq 1$

$$\log \frac{W_{t+1}}{W_t} \leq \frac{1}{1-\beta} \sum_{k=1}^{K} p_t(k)(e-2)\eta^2 \hat{X}_t(k)^2$$

so, we have

$$\log \frac{W_{n+1}}{W_1} \leq \frac{1}{1-\beta} \sum_{t=1}^{n} \sum_{k=1}^{K} p_t(k)(\eta \hat{X}_t(k) + (e-2)\eta^2 \hat{X}_t(k)^2)$$

For any $k = 1, ..., K$, we also have

$$\log \frac{W_{n+1}}{W_1} = \log \sum_{k=1}^{K} e^{\eta \sum_{t=1}^{n} \hat{X}_t(k)} - \log K \geq \eta \sum_{t=1}^{n} \hat{X}_t(k) - \log K$$

By taking the expectations, for any $k = 1, ..., K$, we may write

$$\mathbb{E}\left[(1-\beta)\sum_{t=1}^{n} \hat{X}_t(k) - \sum_{t=1}^{n}\sum_{i=1}^{K} p_t(i)\hat{X}_t(i)\right] \leq (1-\beta)\frac{\log K}{\eta} + (e-2)\eta\,\mathbb{E}\left[\sum_{t=1}^{n}\sum_{k=1}^{K} p_t(k)\hat{X}_t(k)^2\right]$$

$$\sum_{t=1}^{n} X_t(k) - \mathbb{E}[\sum_{t=1}^{n} X_t(I_t)] \leq \beta n + \frac{\log K}{\eta} + (e-2)\eta n K$$

$$\mathbb{E}[R_n(k)] \leq \frac{\log K}{\eta} + (e-1)\eta n K$$

□

References

ECE-GY 9223 Reinforcement Learning Class notes, Prof. Quanyan Zhu

Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems By S prime ebastien Bubeck and Nicolo Cesa-Bianchi

http://chercheurs.lille.inria.fr/∼ghavamza/RL-EC-Lille/Lecture%20Bandit.pdf