# Machine Learning (AIMS) - Week 5, Michaelmas Term 2018
# Lecture : Multi-armed Bandits

Lecturer: Varun Kanade

Scribed by Shaan Desai

These notes are a rough draft. Many thanks to Shaan Desai for scribing. A lightly edited and slightly more proof-read set will be posted in due course.

## 1 Introduction

Set of states and a decision maker
E.g.
state is current state of the video game, agent takes an action and you move to a new state and you get a reward.
Agent takes action $a_t$ at time t
and gets reward $r_t$
depending on action, you can go to any random state (if stochastic).
Underlying assumption of Markovian.
Goal: Maximize cumulative reward!
MAB is simplest form, there are no movements form state to state. Trivial system where you have single state. Rewards don't depend on past actions.

### 1.1 Bandits Overview

Bandits with IID rewards
We have a time horizon T ate each round t: 1. agent picks an action/arm $a_t \in A$

$$|A| = k$$

2. receive and observe reward $r_t \in [0, 1]$

### 1.2 Stochastic vs adversarial

stochastic - IID draw from distribution across time for each arm adversarial - gives us a reward (what does it mean to maximize in this setting)
Maximize the overall reward across the time horizon - make IID assumption that for all $a \in A$ there is a distribution $D_a$ over $[0, 1]$ and $r_{t,a} \sim D_a$ (independent across time)
Action a, $\mu_a$ bernoulli reward model in this case, as long as support is bounded. Even if support is not bounded (sub gaussian) low probability mass away from mean. Good concentration of averages is wanted.
Can generalize to more reward distributions.

### 1.3 applications

Well studied problem, e.g. drug trials - every time you get a patient with disease you want to choose the best treatment as soon as possible. e.g online predictions - maximize revenue by ad clicks. choose many ads and clicks are pulls of the arm.

### 1.4 Notation

Action - $A$ time - $t$, $T$ # of arms - $K$ # rounds - $T$ mean reward of action $a$ - $\mu_a$ best mean reward - $\mu^*$ $\delta(a) = \mu^* - \mu(a)$ reward gap (suboptimality of arm $a$) $a^*$ any optimal action Cumulative Regret

$$R(t) = \mu^* t - \sum_{s=1}^{t} \mu_{a_s}$$

regret is always positive, we dont want it to grow as a function of t. Want the average regret to grow slowly:

$$R(t)/t \to 0$$

as

$$t \to \inf$$

$$max_{a \in A} \sum r_{t,a} - \sum r_{t,a_t}$$

LHS - best action in hindsight RHS - actual reward accumulated
in adversarial it is hard to know which action is best
in stochastic setting:

## 2 Uniform exploration

want to find the best arm which involves exploration, spend too much time exploring - more danger accumulating regret! trade off between the two.

### 2.1 Algorithm

Fix param N
Pull all arms N times $\hat{\mu} = 1/N \sum r_{t,a}$
pick $\bar{a} = argmax \hat{\mu}$
play $\bar{a}$ for remaining T - KN steps

### 2.2 Hoeffding bound/Proof

Let $X_1, ...X_N$ be IID RVs taking values in [0,1] with $E[X_1] = \mu$
then

$$P(|\frac{1}{n} \sum x_i - \mu| > t) \leq 2e^{-2nt^2}$$

Union Bound: $P(AUB) < P(A) + P(B)$
Arm a: $\epsilon_a$ the event that after N draws:

$$|\hat{\mu_a} - \mu_a| \geq \sqrt{\frac{2logT}{N}}$$

$$\epsilon = U\epsilon_a$$

$$P(\epsilon_a) \leq 2e^{-2NlogT/N} = 2e^{-4logT} = 2/T^4$$

$$P(\epsilon) \leq 2K/T^4$$

expected regret is very small. Condition on E not occuring. Suppose algo picks

$$\bar{a}! = a^*$$

$$\hat{\mu a} \geq \hat{\mu_{a^*}}$$

some algebra:

$$\hat{\mu a} \leq \hat{\mu_a^*} + 2\sqrt{\frac{2logT}{N}}$$

also:

$$\mu_{\bar{a}} \geq \mu_a^* - 2\sqrt{\frac{2logT}{N}}$$

$$\Delta(\bar{a}) \leq 2\sqrt{\frac{2logT}{n}}$$

Regret : $\sum \Delta(a_t)$ total regret $\leq KN + 2T\sqrt{\frac{2logT}{N}}$ KN is exploration phase, just accumulate lots of regret. Then only exploit! Need to choose the right N,

$$N = (\frac{2T\sqrt{2logT}}{K})^{2/3}$$

can substitute above into regret above.
total regret $\leq 2K^{1/3}T^{2/3}logT^{1/3}$
quick notes from questions:
$\bar{a}$ : assumed maximum of $\hat{\mu_a}$ probability that E doesn't happen is at most 1 E(regret) = P (regret—event happens)(event happens) + P(regret—doesnt happen)P(doesnt happen)
Issues with algo: stop exploring, need to know how to pull each arm I don't know how long i'm going to play one way is to pretend what T is, double estimate of T after each run

## 3 $\epsilon$ greedy

for each t, with probability $\epsilon_t$ pick a uniform random action w probability 1 -$\epsilon_t$ pick the best action so far
explore in the beginning and let it decay with time

$$\epsilon_t \sim t^{-1/3}$$

$$E[Regret] = O(T^{2/3}(KlogT)^{1/3})$$

can we improve on this?
lets not just pick a random action, might want to be smarter.
Adaptive exploration algorithms
Exploration in phases - successive elimination
we have:

$$\hat{\mu_a}$$

lies in an interval
length of interval:

$$\sqrt{\frac{logT}{N_a}}$$

overlap of mu's determines whether we drop the arm or not
If for arm a, $\exists$ arm a' s.t:

$$\hat{\mu_a} + \sqrt{\frac{2logT}{N_a}} \leq \hat{\mu_a'} - \sqrt{\frac{2logT}{N_a'}}$$

then eliminate a
run through all the arms, cycle through arms and keep eliminating
Regret = $O(\sqrt{KTlogT})$

## 4 UCB - upper confidence bound

for each t,
pick $a \in argmax(\hat{\mu_a}(t) + r_t(a))$
if a is played $N_{a,t}$ times before time t

$$\hat{\mu_a}(t) = 1/N_t(a) \sum r_{a,s_i}$$

$$r_t(a) = \sqrt{\frac{2logT}{N_t(a)}}$$

be optimistic
proving that UCB works!

$$P(|\hat{\mu_t}(a) - \mu_a| \geq r_t(a)) \leq 2/T^4$$

$$r_t(a) = \sqrt{2logT/N_t(a)}$$

rewards table across time $r_{t,j}$
$N_t(a) \leq T$
Condition on this!

$$P(|\hat{\mu_t}(a) - \mu_a| \geq r_t(a)|N_t(a) = K) \leq 2/T^4$$

$p(\epsilon) \leq 2K/T^3$
Fix some time t Algorithm picks some $a_t$

$$\hat{\mu_t}(a_t) + r_t(a_t) \geq \hat{\mu_t}(a^*) + r_t(a^*)$$

UCB

$$\hat{\mu_t}(a_t) + r_t(a_t) \leq \mu(a_t) + 2r_t(a_t)$$

mu hat is lowest value of true

$$\hat{\mu_t}(a^*) + r_t(a^*) \geq \mu(a^*)$$

$$Delta(a_t) = \mu(a^*) - \mu(a_t) \leq 2r_t(a_t)$$

Regret = $\sum_{t=1}^{T} \Delta(a_t) = \sum_{a=1}^{K} \sum_{s=1}^{N_T(a)} 2\sqrt{\frac{2logT}{S}}$
I know:

$$\sum_{a=1}^{K} N_T(a) = T$$

$\leq 2 \sum_{a=1}^{K} \sqrt{2logTN_T(a)}$
on the assumption that $\sum 1/sqrti \sim k\sqrt{n}$
use concavity

$$\leq 200\sqrt{2log(T)KT}$$

sqrt KT is a limitation, this is the holy grail.