Multiarmed bandits:

- Bubecki notes
- Lattimore & Szepesvari

Introduced by William Thompson - medical trials
- was against running a trial blindly, without adapting treatment on the fly, depending upon the efficacy of the drug.

- Suitable in the context of decision making with uncertainty.

- Tech companies use such algorithms for configuring web interfaces for recommendation, pricing.

Classical dilemma.

Round	1	2	3	4	5	6	7	8	9	10
Left	0		10	0		0				10
Right		10			0		0	0	0	

- Left arm seems better - average payoff is 4.
- If you have 10 more pulls, what do you do?

Language of bandits:

Game b/w learner & environment
- played over n rounds, called horizon

In each round t learner chooses an action A_t and gets a reward $X_t \in \mathbb{R}$;

- [Actions called arms. k-armed bandits, have k actions]

A_t depends upon history
$$H_{t-1} = (A_1, X_1, A_2, X_2, \ldots, A_{t-1}, X_{t-1})$$

Policy: $\pi_t : (A \times X)^{t-1} \longrightarrow A.$

· **Environment:** $E_t : (A_1, X_1, \ldots, A_{t-1}, X_{t-1}, A_t) \longrightarrow \mathbb{R}$

mapping from histories ending in actions to reward.

reward on action A_t.

Objective. maximize $\sum X_t$

CHALLENGE: Learner has no idea of environment except that it belongs to an environment class.

· **Evaluation?** Regret:

Definition: The regret of the learner relative to a policy π is the difference b/w the total expected reward using policy π for n rounds and the total expected reward collected by the learner over n rounds;

Regret relative to a set of policies Π, is $\max_{\pi \in \Pi}$ (regret relative to π)

- $\Pi \leftarrow$ Competitor class.

Usually Π is large enough to include the optimal policy for all environments in $\mathcal{E}$.

Example: Suppose $\mathcal{A} = \{1, 2, \ldots, k\}$; An environment is called stochastic Bernoulli if the reward $X_t \in \{0, 1\}$, is binary valued and $\exists\ \mu \in [0,1]^k$, s.t $\Pr[X_t = 1 \mid A_t = a] = \mu_a$.

If you knew the mean vector μ_a, associated to the environment, the optimal policy is the fixed action, $a^* = \underset{a \in [n]}{\arg\max}\, \mu_a$.

. <u>Competitor class</u>: $\Pi = \{\pi_1, \ldots, \pi_k\}$, π_i = play i all the time;

<u>Regret over n rounds</u>:

$$R_n = n \max_{a \in A} \mu_a - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right].$$

. Suppose the learner fixes a policy; If the competitor class is also fixed, the regret depends upon the environment.

Ideal

Regret is small $\forall$ environments

WORST case.

max regret over all environments.

- Main question:

Growth rate of regret as a function on n.

Good learners: $\lim_{n \to \infty} \frac{R_n}{n} \to 0$;

finer questions; Is R_n $O(\sqrt{n})$, $O(\log(n))$

Lower bounds;

- Large environment class –
  Large competitor classes – regret can be demanding;

Care needed in choosing these sets so that

a) Regret guarantees are meaningful.

b) $\exists$ policies which make regret small.

-

FRAMEWORK:

General enough to model anything using a rich environment class;

But then difficult to say much.

So restrict attention to certain kinds of environment classes and competitor classes.

Ex: STOCHASTIC STATIONARY BANDITS.

Environment is restricted to generate rewards in response to each action from a distribution that is specific to that action (and independent of previous action choices & rewards)

Stochastic Gaussian bandits;

- If the action set is $A \in \mathbb{R}^1$, the mean reward for choosing $a \in A$ could follow a

Linear model.

$$X_t = \langle a, \theta \rangle + \eta_t, \quad \theta \in \mathbb{R}^d$$

η_t - standard Gaussian.

In the above example, θ is unknown, and $\mathcal{E} = \mathbb{R}^d$.

Q: Assuming rewards are stochastic - is it reasonable? Too restrictive?
- Isn't the world deterministic?
- What if stochastic assumption does not hold?

In such a scenario how will algorithms perform?

- DROP ALL ASSUMPTIONS on how rewards are generated, except that they lie in a bounded set and are chosen without knowledge of the learner's actions.

ADVERSARIAL BANDITS?

- Needle in a haystack?

TRICK: RESTRICT COMPETITOR CLASSES.

APPLICATIONS:

① A/B testing:

Placing the "Buy it now" button on top right or bottom left?

Previously - commit to a trial of each version by splitting users into 2 groups. Each group sees one version; statistics collected & decision made.

- Problem: NOT ADAPTIVE. Maybe better to stop the trial earlier.
- Can pose as a bandit problem

- Each time t a user enters, a bandit algorithm selects an action $A_t \in \mathcal{A} = \{\text{TOP RIGHT, BOTTOM LEFT}\}$ and $X_t = 1$ if the user purchases the product.

(2) ADVERT PLACEMENT:

- Each round – when a user visits the website.

$\mathcal{A}$ = set of adverts,

Choose $A_t \in \mathcal{A}$, if user clicks $X_t = 1$

- May work for some websites, But this will not be able to target advertisements. – ROCK CLIMBING SHOES / HARNESS

Can incorporate this - information about a user - "context";

Can cluster users and use a separate bandit algorithm for each cluster;

- The need to tailor the solution to your needs.
  # clicks may not be the correct metric.

③ Recommendation Systems:

- Which movies to place in "Browse"?

- Reward measured as a function of whether or not you watched / Rating was good;

- Actions - Movies - set of actions is combinatorially large.
- Each user watches few films. Low rank

matrix factorization.

Problem: Not offline, the learning algorithm has to choose what users see and this in turn affects data.

- If few users are recommended "Pather Panchali", few will watch it and data on this film will be scarce.

④ NETWORK ROUTING:

- Learner learns to direct internet traffic.
- Action - set of paths from source to destination;
- Reward - — time taken for packet to reach.

THEORETICAL ANALYSIS:

- $$R_n = \max_{i=1,\dots K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}$$

Competitor class $= \{1,\dots,K\}$; Learner / forecaster;

If there is stochasticity:

Expected Regret:

$$\mathbb{E} R_n = \mathbb{E}\left[\max_{i=1,\dots K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right]$$

PSEUDO-REGRET:

$$\overline{R}_n = \max_{i=1,\dots K} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum X_{I_t,t}\right]$$

Compares with the optimal action in expectation;

$$\overline{R}_n \le \mathbb{E} R_n.$$

STOCHASTIC BANDIT PROBLEM:

Input: K # arms; Horizon n
unknown: K distribution $\nu_1, \dots, \nu_K$ on $[0,1]$

for each round $t = 1, \dots, n$

(1) Learner chooses $I_t \in \{1, \dots, K\}$

(2) Given I_t, environment draw reward $X_{I_t, t} \sim \nu_{I_t}$ and reveals to learner;

- Set $\mu_i := E[\nu_i]$;

$$\mu^* = \max_{i=1,\dots,K} \mu_i \quad , \quad i^* = \operatorname{argmax}_{i=1,\dots,K} \mu_i$$

for fixed i:

$$E\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t, t}\right] =$$

$$= n\mu_i - \mathbb{E}\left[\sum_{t=1}^{n} X_{I_t, t}\right]$$

Let P_a be the distribution of the ith arm;

$$\mu_a = \int_{-\infty}^{\infty} x \, dP_a(x)$$

density

$\nu = (P_a : a \in A)$

Let $\Delta_a(\nu) = \mu^*(\nu) - \mu_a(\nu)$.

Suboptimality gap of action a.

Let $T_a(t) = \sum_{s=1}^{t} \mathbb{1}\{A_s = a\}$

action in time $s = a$

times a chosen in the first t rounds,

Clearly $T_a(n)$ is a r.v.

Lemma: Regret decomposition lemma:

For any policy π & environment ν, with $\mathcal{A}$ finite or countable and horizon $n \in \mathbb{N}$

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \, \mathbb{E}[T_a(n)].$$

(ie) to keep pseudo-regret down, the learner should try to minimize the weighted sum of expected action counts, weights being $(\Delta_a)_{a \in \mathcal{A}}$ - the suboptimality gap.

Pf: For a fixed t, $\sum_{a \in \mathcal{A}} \mathbb{1}\{A_t = a\} = 1$.

$$\therefore S_n = \sum_t X_t = \sum_t \sum_a X_t \mathbb{1}\{A_t = a\}.$$

$$\therefore \bar{R}_n = n\mu^* - \mathbb{E}[S_n]$$

$$= n\mu^* - \mathbb{E} \sum_{t=1}^{n} \sum_a X_t \mathbb{1}\{A_t = a\}$$

$$= \sum_{a \in A} \sum_{t=1}^{n} \mathbb{E}\left[(\mu^* - X_t)\,\mathbb{1}\{A_t = a\}\right]$$

The expected reward in round t conditioned on A_t is μ_{A_t}.

$$\therefore \quad \mathbb{E}\left[(\mu^* - X_t)\,\mathbb{1}\{A_t = a\} \,\middle|\, A_t\right]$$

$$= \mathbb{1}\{A_t = a\}\,\mathbb{E}\left[(\mu^* - X_t) \,\middle|\, A_t\right]$$

$$= \mathbb{1}\{A_t = a\}\,(\mu^* - \mu_{A_t})$$

$$= \mathbb{1}\{A_t = a\}\,(\mu^* - \mu_a)$$

$$= \mathbb{1}\{A_t = a\}\,\Delta_a.$$

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}(X|Y)]$$

$$\therefore \sum_{a \in A} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{E}\left[(\mu^* - X_t)\,\mathbb{1}\{A_t = a\} \,\middle|\, A_t\right]\right]$$

$$= \sum_{a} \sum_{t=1}^{n} \mathbb{E}\left[\mathbb{1}\{A_t = a\}\,\Delta_a\right]$$

$$= \sum_{a} \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{1}\{A_t = a\}\,\Delta_a\right]$$

$$= \sum_a \Delta_a \, \mathbb{E}\Big(\underbrace{\sum_{t=1}^{n} \mathbb{1}(A_t = a)}_{} \Big)$$

$$= \sum_a \Delta_a \, \mathbb{E}\big(T_a(n)\big).$$