# Framework and Regret.

Ref: Chap 4, Bandit Algorithms.

1. Basic setting — unstructured / structured
2. Regret — frequentist / Bayesian

3. Regret decomposition.

Notation: 1. $\mathcal{A}$ is the set of arms (aka actions)

a. Choose any $a \in \mathcal{A}$
b. Receive reward drawn from an (unknown) distribution $P_a$.

2. $\nu$ = joint distribution across arms.

e.g.: arms are independent $\Rightarrow$ $\nu$ is a product measure

$= P_{a_1} \times P_{a_2} \times \cdots$ if arms are discrete.

i.e., $\nu = (P_a, a \in \mathcal{A})$ notation for product dist.

| Name | Symbol | Definition |
|---|---|---|
| Bernoulli | $\mathcal{E}^k_{\mathcal{B}}$ | $\{(\mathcal{B}(\mu_i))_i : \mu \in [0,1]^k\}$ |
| Uniform | $\mathcal{E}^k_{\mathcal{U}}$ | $\{(\mathcal{U}(a_i,b_i))_i : a,b \in \mathbb{R}^k$ with $a_i \le b_i$ for all $i\}$ |
| Gaussian (known var.) | $\mathcal{E}^k_{\mathcal{N}}(\sigma^2)$ | $\{(\mathcal{N}(\mu_i,\sigma^2))_i : \mu \in \mathbb{R}^k\}$ |
| Gaussian (unknown var.) | $\mathcal{E}^k_{\mathcal{N}}$ | $\{(\mathcal{N}(\mu_i,\sigma_i^2))_i : \mu \in \mathbb{R}^k$ and $\sigma^2 \in [0,\infty)^k\}$ |
| Finite variance | $\mathcal{E}^k_{\mathbb{V}}(\sigma^2)$ | $\{(P_i)_i : \mathbb{V}_{X\sim P_i}[X] \le \sigma^2$ for all $i\}$ |
| Finite kurtosis | $\mathcal{E}^k_{\mathrm{Kurt}}(\kappa)$ | $\{(P_i)_i : \mathrm{Kurt}_{X\sim P_i}[X] \le \kappa$ for all $i\}$ |
| Bounded support | $\mathcal{E}^k_{[a,b]}$ | $\{(P_i)_i : \mathrm{Supp}(P_i) \subseteq [a,b]\}$ |
| Subgaussian | $\mathcal{E}^k_{\mathrm{SG}}(\sigma^2)$ | $\{(P_i)_i : P_i$ is $\sigma$-subgaussian for all $i\}$ |

**Table 4.1** Typical environment classes for stochastic bandits. $\mathrm{Supp}(P)$ is the (topological) support of distribution $P$. The kurtosis of a random variable $X$ is a measure of its tail behavior and is defined by $\mathbb{E}[(X-\mathbb{E}[X])^4]/\mathbb{V}[X]^2$. Subgaussian distributions have similar properties to the Gaussian and will be defined in Chapter 5.

Bandit Algorithms, Lattimore and Szepesvari, 2019 (pp. 58)

3. $\mathcal{E}$ = environment = set of possible distributions

i.e., $\nu \in \mathcal{E}$

**Stochastic Bandit Problem:**

* nature/adversary chooses a fixed $\nu \in \mathcal{E}$

→ this is unknown to the player.

* At each (discrete) time $t$, player chooses

arm $A_t \in \mathcal{A}$. Nature then uses predetermined $\nu$ to determine arm dist $P_{A_t}$, and draws a sample $X_t \sim P_{A_t}$

* $X_t$ = reward at time $t$

* Player thus observes $(A_t, X_t)$ at time $t$.

* Player uses all causal information:

$$\{(A_s, X_s), s = 1, 2, \dots, t\}$$

to determine action $A_{t+1} \in \mathcal{A}$.

Two settings:

1. Unstructured environment $\mathcal{E}$: $\nu$ is a product measure, i.e., arms do not provide any information about each other.

$$\nu = (P_a, a \in \mathcal{A}) = P_{a_1} \times P_{a_2} \cdots \quad \text{(discrete arms)}$$

(2) Structured environment $\mathcal{E}$: arms "leak" information.

e.g.: linear bandits: $\theta \in \mathbb{R}^d$ fixed but unknown

$$\mathcal{E} = \{ \nu_\theta \ , \ \theta \in \mathbb{R}^d \}$$

where $\nu_\theta = \{ P_{a,\theta} \ , \ a \in \mathcal{A} \}$,

with $P_{a,\theta} = N(\langle a, \theta \rangle, \sigma^2)$

or $\text{Bernoulli}(\langle a, \theta \rangle)$

$\vdots$

Here, given $\theta$, arm distributions are known for all arms.

Since $\theta \in \mathbb{R}^d$, there are efficient ways to estimate $\theta$ from $\tilde{O}(d)$ samples under suitable assumptions.

$\boxed{\text{Regret}}$: The performace metric.

① $\boxed{\text{Frequentist viewpoint:}}$ $\nu \in \mathcal{E}$ picked by nature.

A genie is given this information. Then, the genie uses this to determine best action/arm.

Define $\{X_t^*, \; t = 1, 2, \cdots\}$ the associated reward for the genie.

e.g.: Finite $K$ arms, unstructured environment, with $\nu = P_{a_1} \times P_{a_2} \cdots \times P_{a_K}$

with $P_{a_i} \sim \text{Bernoulli}(\mu_i)$, $\mu_i \in [0,1)$.

Then, WLOG, let $\mu_1 \geq \mu_2 \geq \mu_3 \cdots$

In this case, genie is given knowledge of $\{\mu_1, \cdots \mu_K\}$ and the arm dist. (Bernoulli)

Thus, $X_t^* \sim \text{Bernoulli}(\mu_1)$.

The <u>PLAYER</u> does not have knowledge of $\mu_i$, $i=1,2,\dots,K$.

Plays action $A_s$ at time $s$, and observes $X_s \sim P_{A_s}$, $s=1,2,\dots,t$.

environment

$$\text{Regret} = R_t(\pi,\nu) = \underbrace{E\left[\sum_{s=1}^{t} X_s^*\right]}_{\text{Genie expected reward until time } t} - \underbrace{E\left[\sum_{s=1}^{t} X_s\right]}_{\text{player expected reward until time } t}$$

time horizon

player policy

<u>e.g.:</u>

Continuing our example,

$$E\left[\sum_{s=1}^{t} X_s^*\right] = \mu_1 t$$

---

<u>Notation</u> : $\mu^*(\nu) = E[X_1^*]$ under genie policy that knows $\nu$.

<u>Sub-optimality gap</u>: $\Delta_a(\nu) = \mu^*(\nu) - \mu_a(\nu)$

where $\mu_a(\nu) = E_{P_a}[Z] \quad Z \sim P_a$

$T_a(t)$ = Number of times arm $a$ has been played until time $t$ under policy $\pi$.

$$= \sum_{s=1}^{t} \chi_{\{A_s = a\}}$$

indicator function

Lemma (Regret Decomposition)

Suppose that $|\mathcal{A}|$ is countable. Then:

$$R_t(\pi, \nu) = \sum_{a \in \mathcal{A}} \Delta_a(\nu) \cdot T_a(t).$$

Bayesian Regret: Average regret when we have some prior knowledge on how $\nu \in \mathcal{E}$ is chosen by the environment.

More specifically, suppose $\nu \in \mathcal{E}$ is chosen using some distribut $Q$ with support over $\mathcal{E}$.

Then, for some player policy $\pi$,

$$BR_t(\pi, Q) = \text{Bayesian regret using prior } Q$$

$$= \mathbb{E}_{\nu \sim Q}\left[ R_t(\pi, \nu) \right]$$

From now on (until towards end of semester), we will focus on FREQUENTIST REGRET.

Notation abuse: We will use $R_t$ and hide the dependence on $(\pi, \nu)$ in most cases