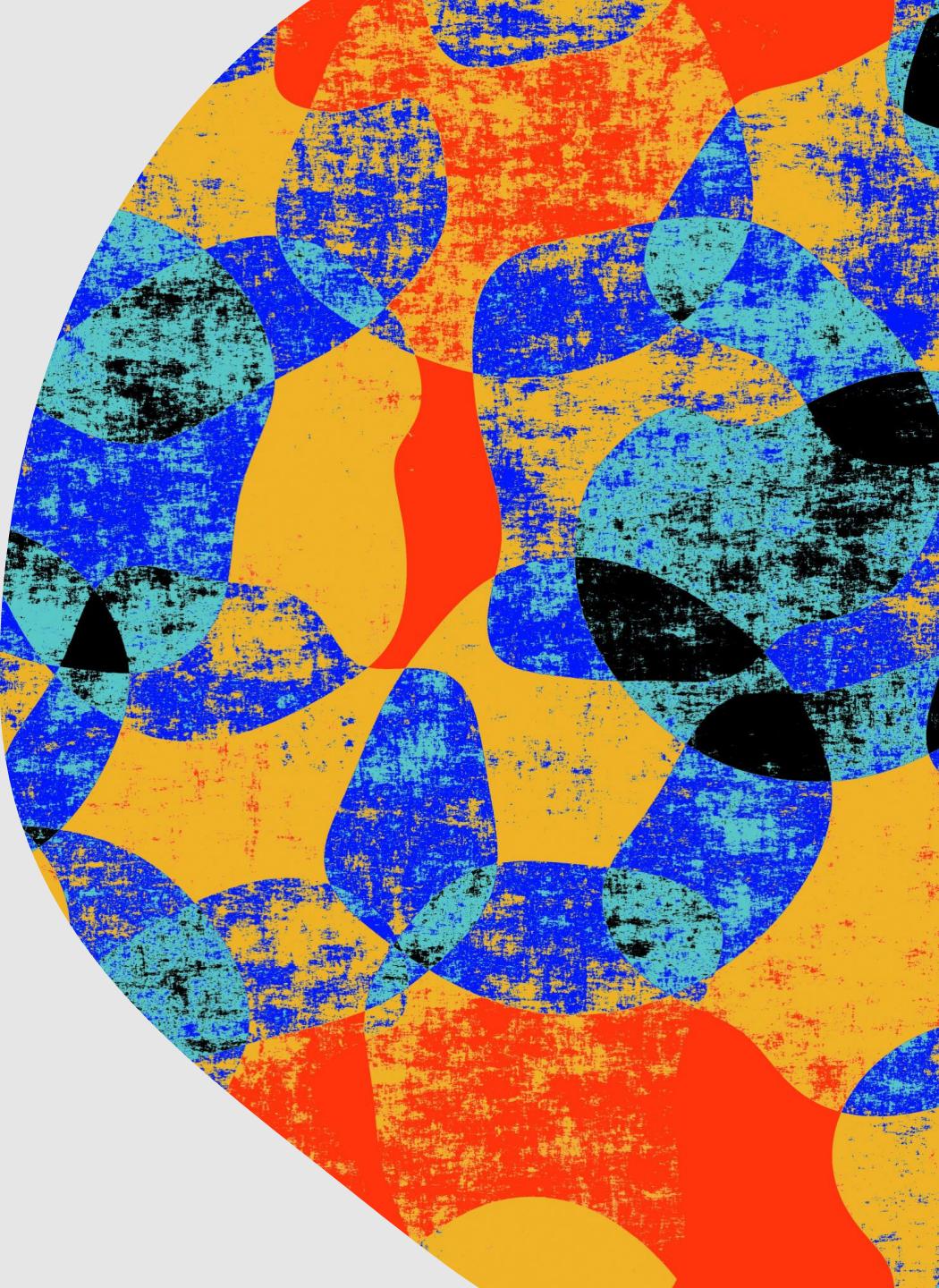


Bandit Algorithms

Chapter 4: Stochastic Bandits

Authors: Tor Lattimore and Csaba Szepesvári

Slides: Hunter Heidenreich



Core Assumptions

$$\nu = (P_a : a \in \mathcal{A})$$

Core Assumptions

Distribution as indexed by selected action

$$\nu = (P_a : a \in \mathcal{A})$$

The diagram illustrates the concept of a stochastic bandit. It features a mathematical expression $\nu = (P_a : a \in \mathcal{A})$. A curly arrow originates from the left side of the expression and points to the letter 'a' in P_a , which is highlighted with a small circle. Another curly arrow originates from the right side of the expression and points to the symbol \mathcal{A} , which is also highlighted with a small circle. Labels provide context: 'A stochastic bandit' is positioned below the expression, 'Distribution as indexed by selected action' is positioned above the letter 'a', and 'Set of actions' is positioned above the symbol \mathcal{A} .

A stochastic bandit

Set of actions

Core Assumptions

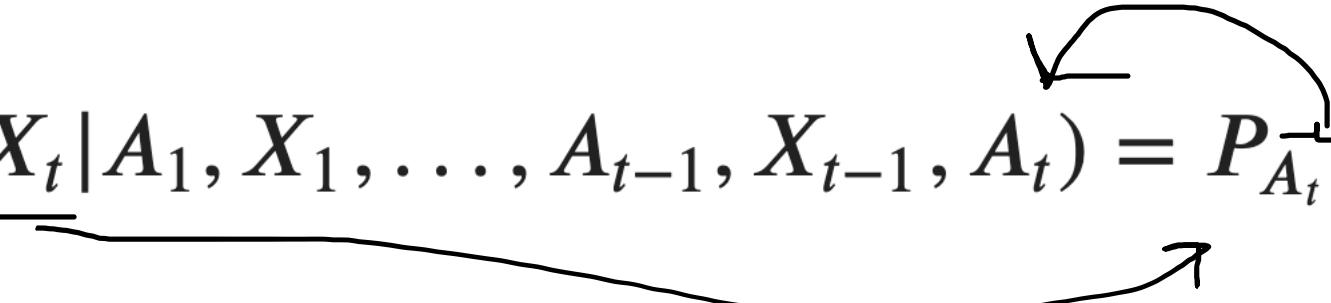


- A learner interacts with the bandit:
 - Time horizon: n
 - Specific time step: t (in $\{1, \dots, n\}$)
 - Specific action selected: A_t
 - Specific reward received: X_t
- This creates a *probability measure* over the sequence:
 - $A_1, X_1, A_2, X_2, \dots, A_n, X_n$

Core Assumptions – Sequence Assumptions

$$p(\underline{X_t} | \underline{A_1}, X_1, \dots, A_{t-1}, X_{t-1}, A_t) = P_{\overline{A_t}}$$

dictated by



$$\pi_t(\cdot | \underline{A_1}, X_1, A_2, X_2, \dots, \underline{A_{t-1}}, \underline{X_{t-1}}).$$

\equiv



No access to the future

The Learning Objective

$$S_n = \sum_{t=1}^n X_t$$

(reward maximization)

??? Why is this not optimization??

- Unknown time horizon
 - (sometimes we know, sometimes we don't)
 - (not a serious issue and is typically easy to handle)
- Cumulative reward is a random quantity
- Unknown distributions for each arm

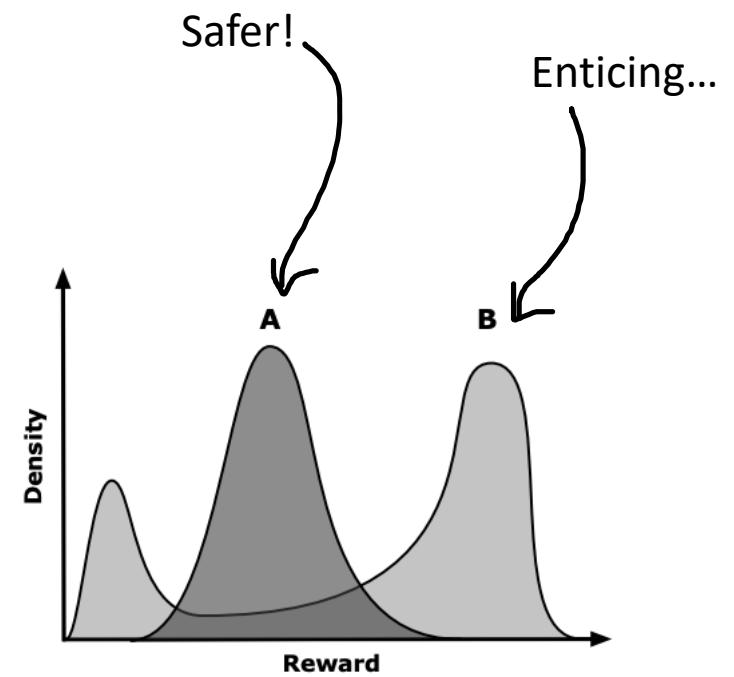


Figure 4.1 Alternative revenue distributions



Environmental Class

What do we know about the class of problems our learner will interact with?

Unstructured

- No action produces information about another action
- More formally:

$$\mathcal{E} = \{\nu = (P_a : a \in \mathcal{A}) : P_a \in \mathcal{M}_a, \forall a \in \mathcal{A}\}$$

Examples of Unstructured Environment Classes

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

Table 4.1 Typical environment classes for stochastic bandits. $\text{Supp}(P)$ is the (topological support of distribution P . The kurtosis of a random variable X is a measure of its tail behaviour 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.

Structured Bandits

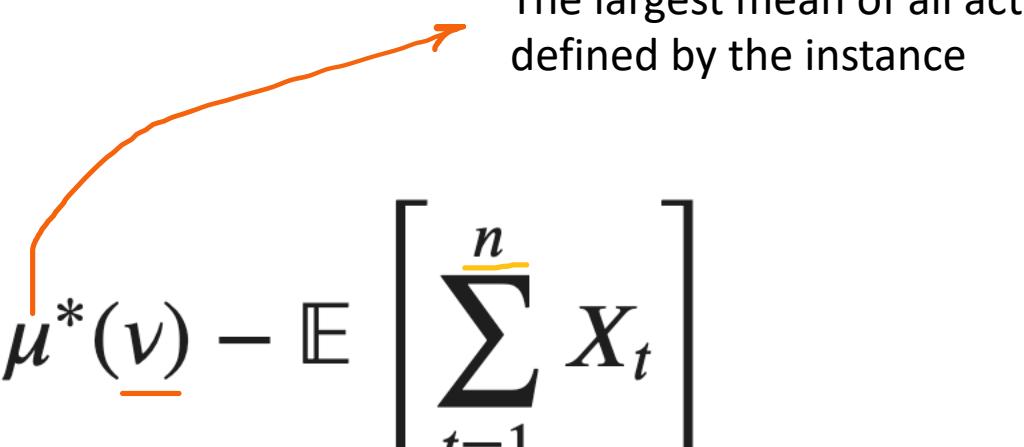
- Actions are correlated in some manner
- Any bandit problem that is not *strictly unstructured*
- A learner can obtain info about an action it never plays!

EXAMPLE 4.1. Let $\mathcal{A} = \{1, 2\}$ and $\mathcal{E} = \{(\mathcal{B}(\theta), \mathcal{B}(1 - \theta)) : \theta \in [0, 1]\}$. In this environment class, the learner does not know the mean of either arm, but can learn the mean of both arms by playing just one. The knowledge of this structure dramatically changes the difficulty of learning in this problem.

Learner Performance – The Regret

Defined for:

- A specific learner policy
- A specific bandit instance
- A specific time horizon


$$R_n(\underline{\pi}, \underline{v}) = \underline{n} \mu^*(\underline{v}) - \mathbb{E} \left[\sum_{t=1}^{\underline{n}} X_t \right]$$

The largest mean of all actions,
defined by the instance

The Regret Lemma

LEMMA 4.4. *Let ν be a stochastic bandit environment. Then,*

(a) $R_n(\pi, \nu) \geq 0$ for all policies π ;

- No negative regret
- It is impossible to better than selecting an optimal action for every round

The Regret Lemma

LEMMA 4.4. *Let ν be a stochastic bandit environment. Then,*

- (a) $R_n(\pi, \nu) \geq 0$ for all policies π ;
- (b) the policy π choosing $A_t \in \operatorname{argmax}_a \mu_a$ for all t satisfies $R_n(\pi, \nu) = 0$;
 - There exists a policy that achieves 0 regret
 - That policy is the policy of selecting an action in the optimal set

The Regret Lemma

LEMMA 4.4. *Let ν be a stochastic bandit environment. Then,*

- (a) $R_n(\pi, \nu) \geq 0$ for all policies π ;
 - (b) the policy π choosing $A_t \in \operatorname{argmax}_a \mu_a$ for all t satisfies $R_n(\pi, \nu) = 0$; and
 - (c) if $R_n(\pi, \nu) = 0$ for some policy π , then $\mathbb{P}(\mu_{A_t} = \mu^*) = 1$ for all $t \in [n]$.
-
- Achieving 0 regret is possible **if and only if** the learner knows the exact bandit that is being faced beforehand

Regret Objectives – Sub-linear

for all $\nu \in \mathcal{E}$,

$$\lim_{n \rightarrow \infty} \frac{R_n(\pi, \nu)}{n} = 0.$$

for all $\nu \in \mathcal{E}$,

$$R_n(\pi, \nu) \leq Cn^p.$$

$p < 1$

Regret Objectives – Factorization

for all $n \in \mathbb{N}$, $\nu \in \mathcal{E}$,

$$R_n(\pi, \nu) \leq C(\nu) \underbrace{f(n)}_{\text{Regret due to instance}}.$$

Regret due to horizon

Decomposing the Regret

LEMMA 4.5 (Regret decomposition lemma). *For any policy π and stochastic bandit environment ν with \mathcal{A} finite or countable and horizon $n \in \mathbb{N}$, the regret R_n of policy π in ν satisfies*

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E} [T_a(n)] . \quad (4.5)$$



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

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