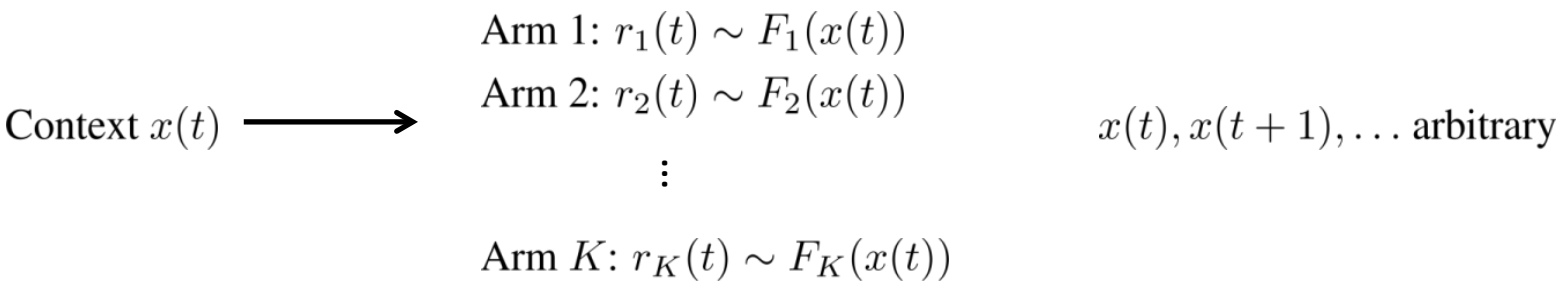
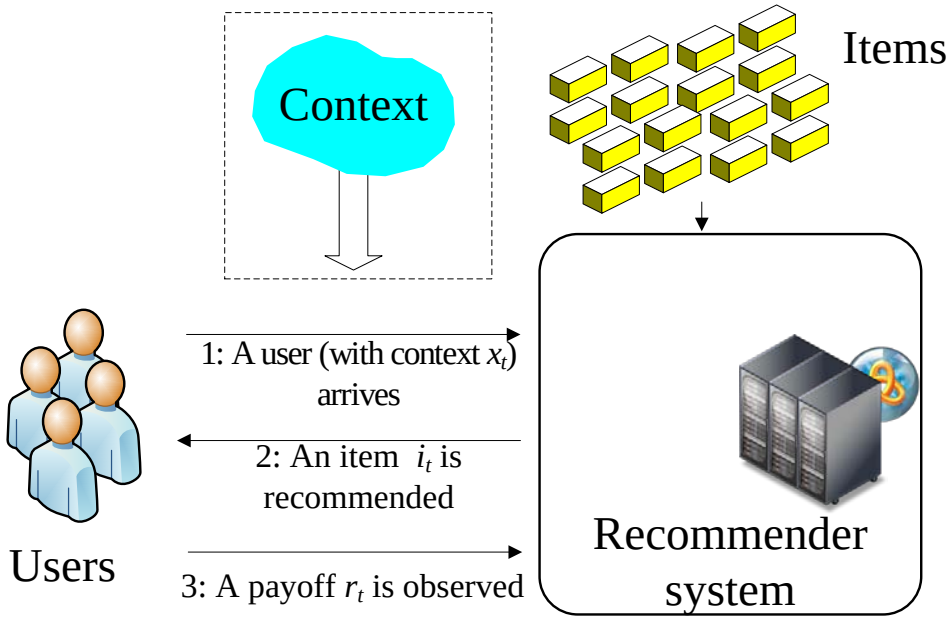
Contextual Multi-armed Bandits

[Slivkins 09] [Lu 2010] [Chu 2011] [Langford2007] [Hazan2007]



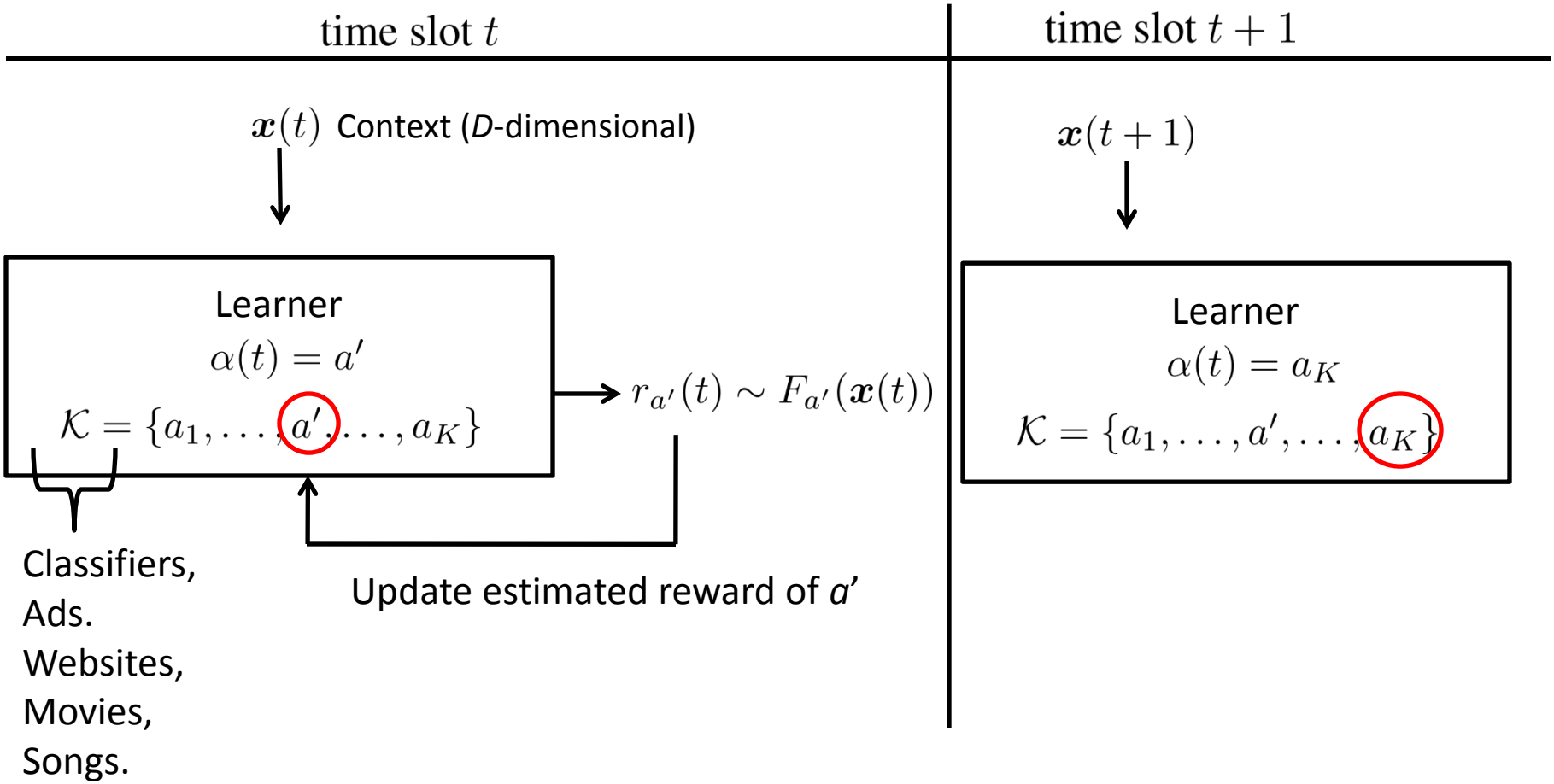
Agent observes the context → **Selects the arm** → **Obtains the reward of the selected arm**

Modeling Recommender Systems as Contextual Bandits



System Model for the Contextual Bandit Problem

•Discrete time slots *t*=1,2,…



Music Dataset

•Randomly pick a user from the dataset for each time *t*.

Contexts

Actions

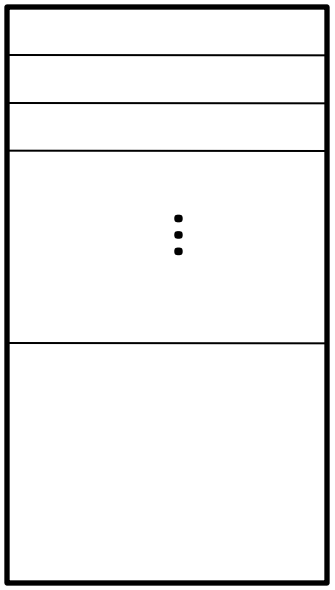
Age	Gender	Freq	When	Activity	Movie	Class	Jazz	Rap	Pop
2	1	2	3	2	3	5	5	2	4

•Observe the contexts, choose an action = Rap

•Observe only rating 2 (rating of the user for Rap) & Update the learning algorithm

•Other ratings of the user are not observed because other actions are not chosen

Music Dataset



You are given:
60% of the dataset (60% of the users)
to design and validate your algorithm

Your algorithms performance will be tested on:
40% of the dataset (40% of the users)
(This portion will only be made available after the algorithm
design phase.)

An Example:
Contextual Learning with Uniform Partition (CUP)

•Learns fast using similarities between contexts.

•Creates a partition of the context space and learns together for each set in the partition.

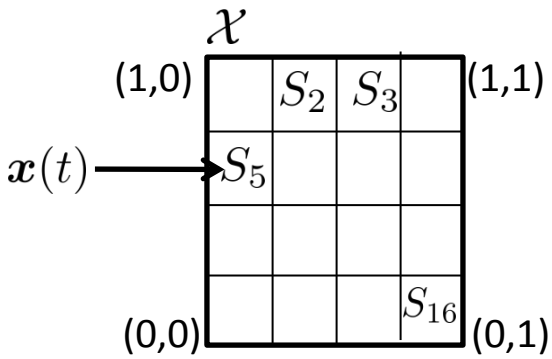
• $\mathcal{P}$ is a partition of $\mathcal{X}$ if $\bigcup_{s\in\mathcal{P}} s = \mathcal{X}$ and $s \cap s' = \emptyset$

Ex: $\mathcal{X} = [0,1], \quad \mathcal{P} = \{[0,1/3), [1/3,1]\}$

CUP – What Happens in a Time Slot

Example: 2 dimensional context space, *D*=2.

Step 1: Find the set in the partition that the context belongs to.



$m_T = 4$ slices per dimension

$\mathcal{P} = \{S_1, S_2, \ldots, S_{16}\}$

Learning is done ***independently*** for each set S_l based only on the ***past history in the set***.

What does CUP keep for each set S_l?

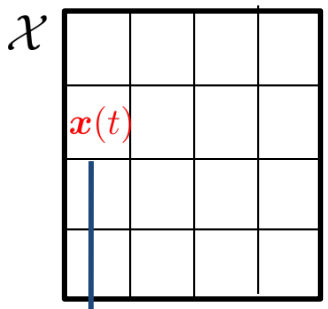
(i) Sample mean reward of arm *i* : $\hat{\pi}_i(S_l)$

(ii) Counter for the number of selections of arm *i* : $N_i^{\exp}(S_l, t)$

(iii) Exploration control function: $F(t) = t^z \log t$

CUP – What Happens in a Time Slot

Step 2: Explore or Exploit



$S(t)$: Set in $\mathcal{P}$ that contains $\boldsymbol{x}(t)$

Ex: $S(t) = S_5$

if any $N_i^{\exp}(S(t), t) \leq F(t)$

Explore arm i

- Get reward $r_i(t)$
- Update i's estimated reward on $S(t)$

$$\hat{\mu}_i(S(t)) = \frac{\hat{\mu}_i(S(t)) \times N_i^{\exp}(S(t),t) + r_i(t)}{N_i^{\exp}(S(t),t)+1}$$

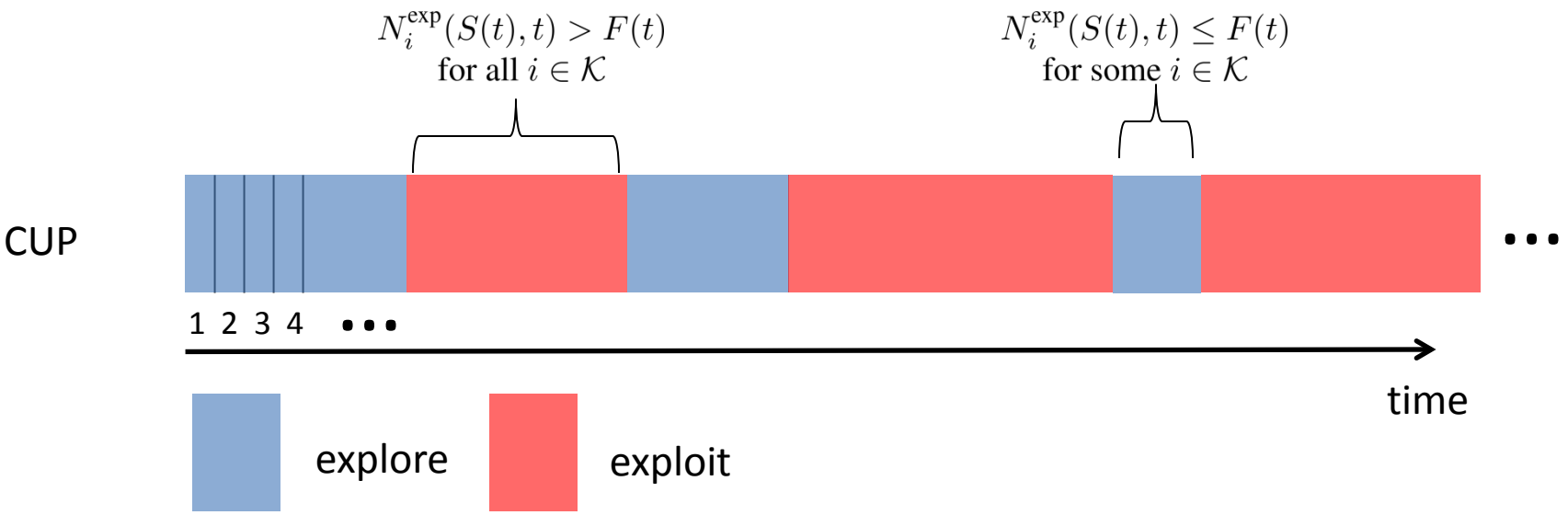
$$N_i^{\exp}(S(t), t+1) = N_i^{\exp}(S(t), t+1) + 1$$

else

Exploit

- select arm $\alpha(t) = \arg\max_{i \in \mathcal{K}} \hat{\mu}_i(S(t))$
- Get reward $r_{\alpha(t)}(t)$
- Update i's estimated reward on $S(t)$

A Sample Path - What Happens Over Time



- When to explore or exploit depends on the context arrival process.
- Exploration rate decreases over time.

Example Results

$F(t) = t^z \log t$ CUP z sweep	1/16	1/8	1/4	1/3	1/2.5	1/2
Exploit %	92.66	88.57	82.70	70.46	57.37	29.44
Average rating (max=1)	0.95	0.92	0.89	0.81	0.73	0.57
Average rating in Exploitations (max=1)	0.996	0.996	0.996	0.997	0.998	0.998