Multi-Armed Bandit

Multi Armed Bandit Problem

Explore vs Exploit

→ try different machine

→ Pull arm on "best" machine

$\mathcal{S}$ = Set of Strategies

$= [n]$

Σ = Set of all possible cost functions

Costs in range $[0, 1]$

$\Sigma = [0, 1]^{[n]}$

Sequence of cost functions

$$c_1, c_2, \ldots, c_T \in \Sigma$$

we pick

x_t for $t \geq 1$

depends only on $\{c_i(x_i)\}_{i=1}^{t-1}$

$$\hat{R}(ALG, T) = \max_{x \in \mathcal{X}} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} c_t(x_t) - c_t(x)\right]$$

Adversary

can pick cost functions ahead of time

OR

can have adaptive Adversary,

can know $x_1 \ldots x_{t-1}$ but $x_1 \ldots x_T$

Ethical Clinical Trials

typically

n = # of treatments

T = # of patients in trial

X_t is the treatment for patient t

$C_t(X_t)$ is how well treatment works

0 = success

1 = Death

Server Selection in networks

We have n servers

x_t is server we pick at DNS Query t

$c_t(x_t)$ is the time it took

Internet Ad Placement

time t we got a visitor

n ads we could show

x_t is the ad that we show them

$$c_t(x_t) = \begin{cases} 1 \text{ if click ad} \\ 0 \text{ if no click} \end{cases}$$

Identical goods for sale

$\mathcal{X}$ = prices you could offer

x_t is the price you offer visitor t

$c_t(x_t) = p$ if they buy
or 0 otherwise

Overlay routing in networks

$S = v_0, v_1, \ldots, v_L = t$

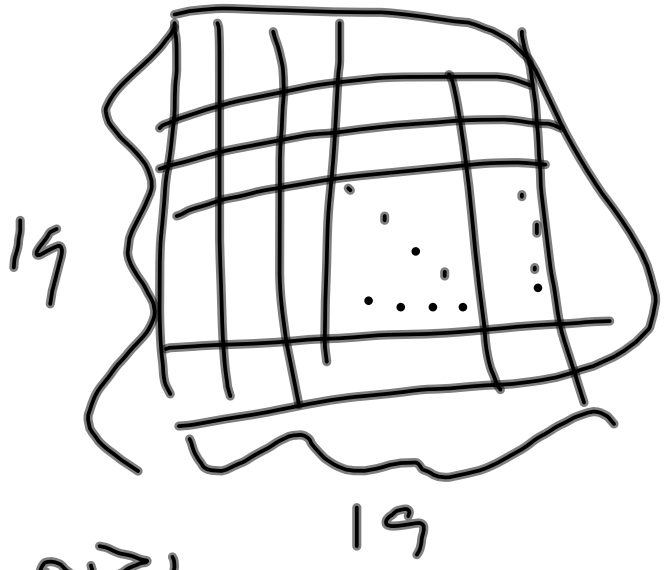
$d(v, w)$

$$\text{total time} = \sum_{i=0}^{L-1} d(v_i, v_{i+1})$$

S = set of paths from s to t

GO

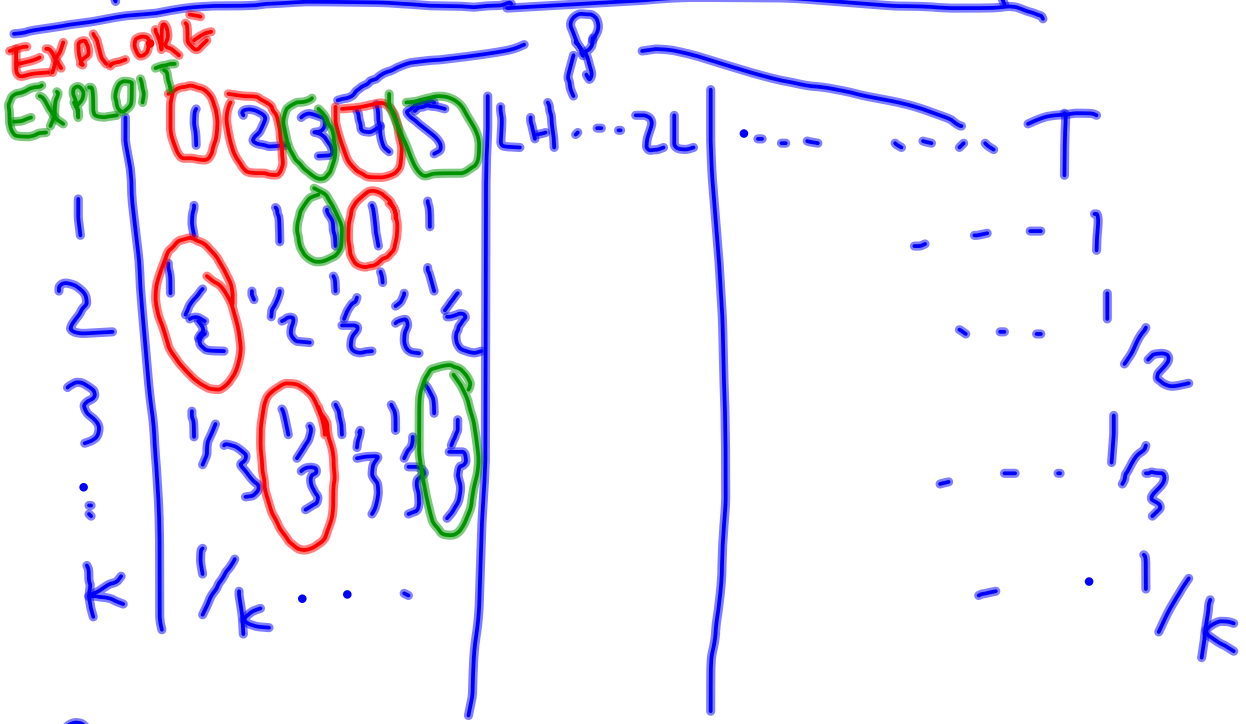
S = moves you could make



X_t = move that you pick

$C_t(x_t)$ = win rate for that move at that time

Phased Simulation



Pick $L > k (= n)$

(Assume $T = P \cdot L$)

BEX depends on parameter $\varepsilon > 0$

Hedge FPL $\Rightarrow$ BEX(ε)

$$\hat{R}(\text{BEX}(\varepsilon), t) \leq \varepsilon t + O\left(\frac{\log n}{\varepsilon t}\right)$$

Function $\tau_j(i) \rightarrow \{jL+1, \ldots, (j+1)L\}$

$\tau_j(i)$ is 1-to-1. chosen randomly.

τ_j represents the time steps in interval j for which we run EXPLORING.

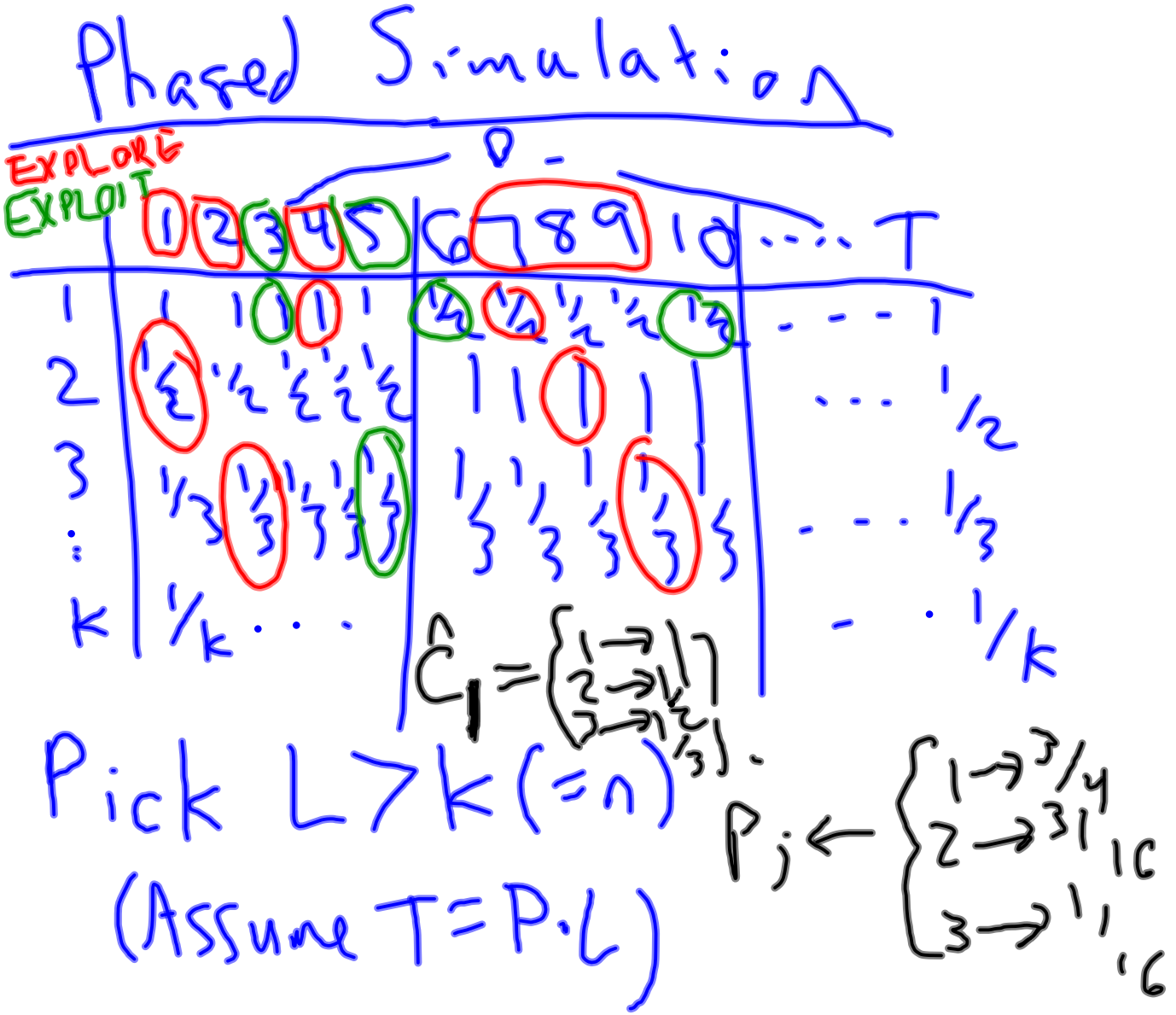
The exploit steps are then the remaining $[T] \setminus$ EXPLORE steps.

$\hat{C}_j$ is func. such that $\hat{C}_j(i) = C_{t_j(i)}(i)$

p_j : dist. over $[n]$ from BEX at time j.

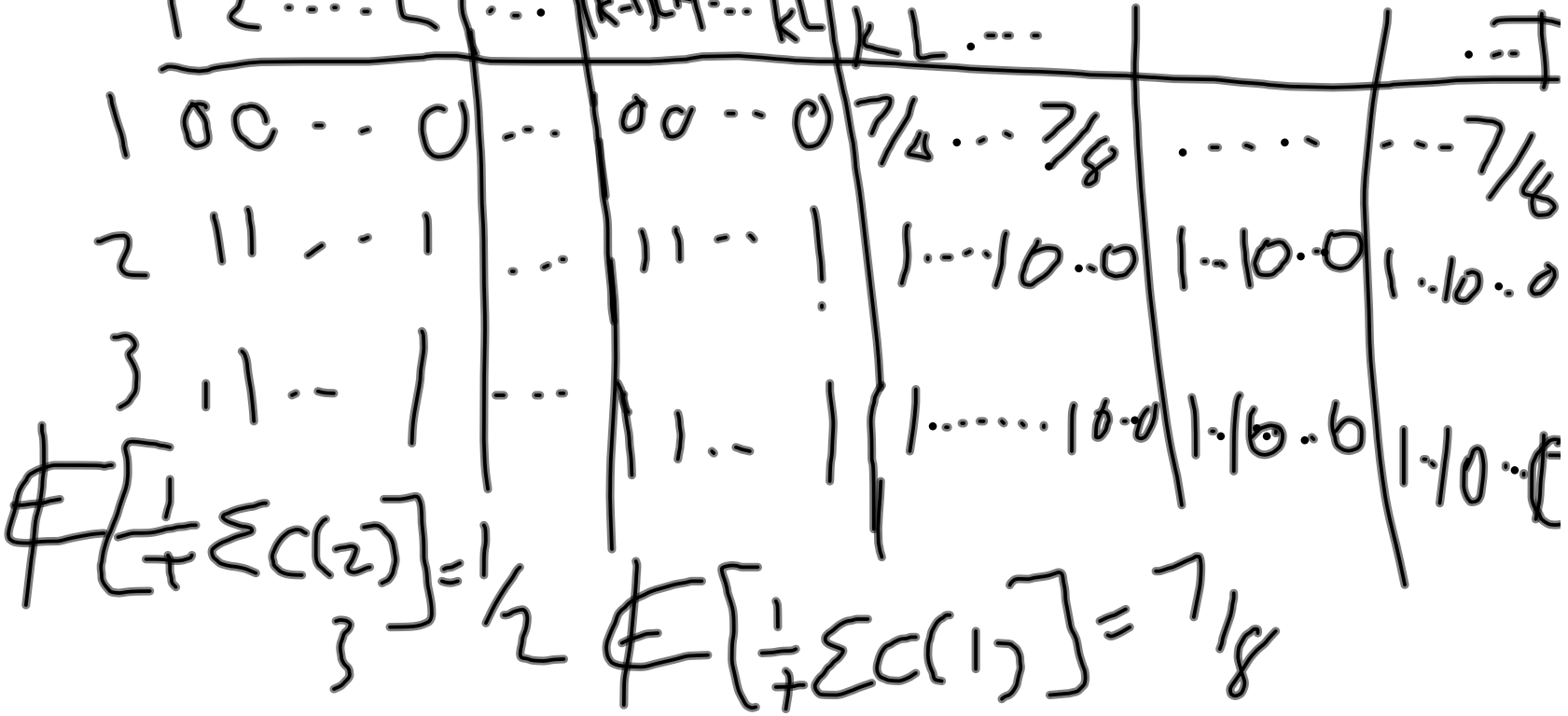
Pick $X_t = \begin{cases} i & \text{if } \tau_j(i) = t \\ \hat{x}_j & \text{otherwise,} \end{cases}$

where $\hat{x}_j \leftarrow p_j$.


Phased Simulation
EXPLORE
EXPLOIT
1 2 3 4 5 6 7 8 9 10 ... T
1
2
3
⋮
k
Pick L > k (= n)
(Assume T = P·L)

Analysis of PSIM = "phased simulation"

$$\frac{1}{P}\sum_j \hat{C}_j \approx \frac{1}{T}\sum_t C_t$$



Assume C_t is not adaptive.

Lemma 1. $\mathbb{E}\left[\hat{c}_j \mid \mathcal{F}_{<j}\right] = c_j$.

$$= \frac{1}{L} \sum_{t \in \Phi_j} c_t.$$

Proof: LHS & RHS : $[n] \to [0,1]$.

Let arb. $i \in [n]$. WTS:

$\mathbb{E}\left[\hat{c}_j(i) \mid \mathcal{F}_{<j}\right] = c_j(i)$. ✓

↳ obvious. kind of. By unif. dist. of τ_j.

Lemma 2. Let $x \in \mathcal{Y}_j$

$$\mathbb{E}\left[\hat{c}_j(\hat{x}_j) \mid \mathcal{F}_{<j}\right] = \mathbb{E}\left[c_j(x_j) \mid \mathcal{F}_{<j}\right]$$

Proof. $\mathbb{E}\left[\hat{c}_j(\hat{x}_j) \mid \mathcal{F}_{<j}\right]$

$$= \mathbb{E}\left[\sum_x p_j(x)\, \hat{c}_j(x) \mid \mathcal{F}_{<j}\right]$$

$$= \sum_x p_j(x)\, \mathbb{E}\left[\hat{c}_j(x) \mid \mathcal{F}_{<j}\right]$$

$$= \mathbb{E}\left[c_j(x_j) \mid \mathcal{F}_{<j}\right].$$

Lemma 3. Let $x \in \mathcal{Y}$.

$$\frac{1}{T}\mathbb{E}\left[\sum_{t \in \text{EXPLOIT}} \left(c_t(y_t) - c_t(x)\right)\right] \leq \xi + \frac{\log n}{\xi p}.$$

Proof.

$$\frac{1}{T}\sum_{t \in \text{EXPLOIT}} \mathbb{E}\left[c_t(x_t) - c_t(x)\right]$$

$$= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[c_t(\hat{x}_j) - c_t(x)\right]$$

$$= \frac{1}{p}\sum_{j=1}^{p} \mathbb{E}\left[\bar{c}_j(\hat{x}_j) - \bar{c}_j(x)\right]$$

$$\leq \varepsilon + \frac{\log n}{\varepsilon p}$$

Lemma 4. Let $x \in \mathcal{Y}$.

Then $\frac{1}{T} \sum_{t \in \text{EXPLORE}} \mathbb{E}\left[c_t(x_t) - c_t(x)\right] \leq \frac{np}{T}$.

Proof. $c_t(x_i) - c_t(x) \leq 1$.

EXPLORE has $n \cdot p$

Theorem 5. Let $\varepsilon = \left(\frac{n \log n}{T}\right)^{1/3}$,

$$p = \log^{1/3}(n)\left(\frac{1}{n}\right)^{2/3}.$$

Then, $\hat{R}(PSim, T) = O\left(\left(\frac{n \log n}{T}\right)^{1/3}\right)$

Proof. $\hat{R}(PSim, T) \leq \varepsilon + \frac{\log n}{\varepsilon p} + n p$.