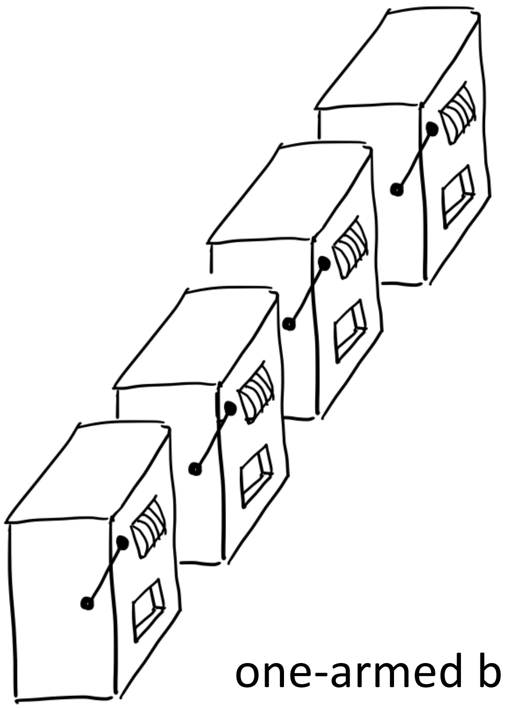
Multi-armed Bandits
Part 1: Basic Algorithms

Cynthia Rudin

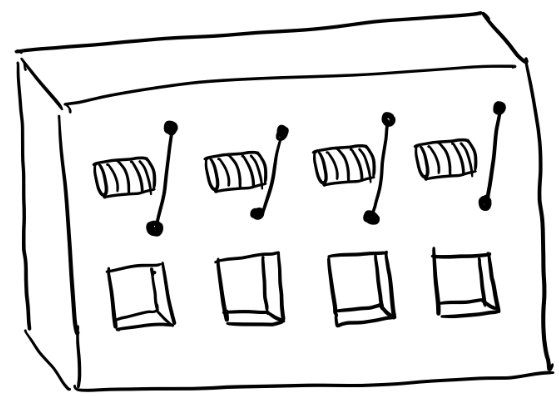
Duke University

Multi-armed bandit

Exploration vs exploitation



one-armed bandits



“multi-armed” bandit

Multi-armed bandit

Applications:

- Ad serving
  - Arms – possible ads
  - Reward – a click
- Website optimization
  - Arms – possible website options
  - Reward – user engagement
- Clinical Trials
  - Arms: possible medications
  - Reward: health outcomes

(Alternative to massive AB testing)

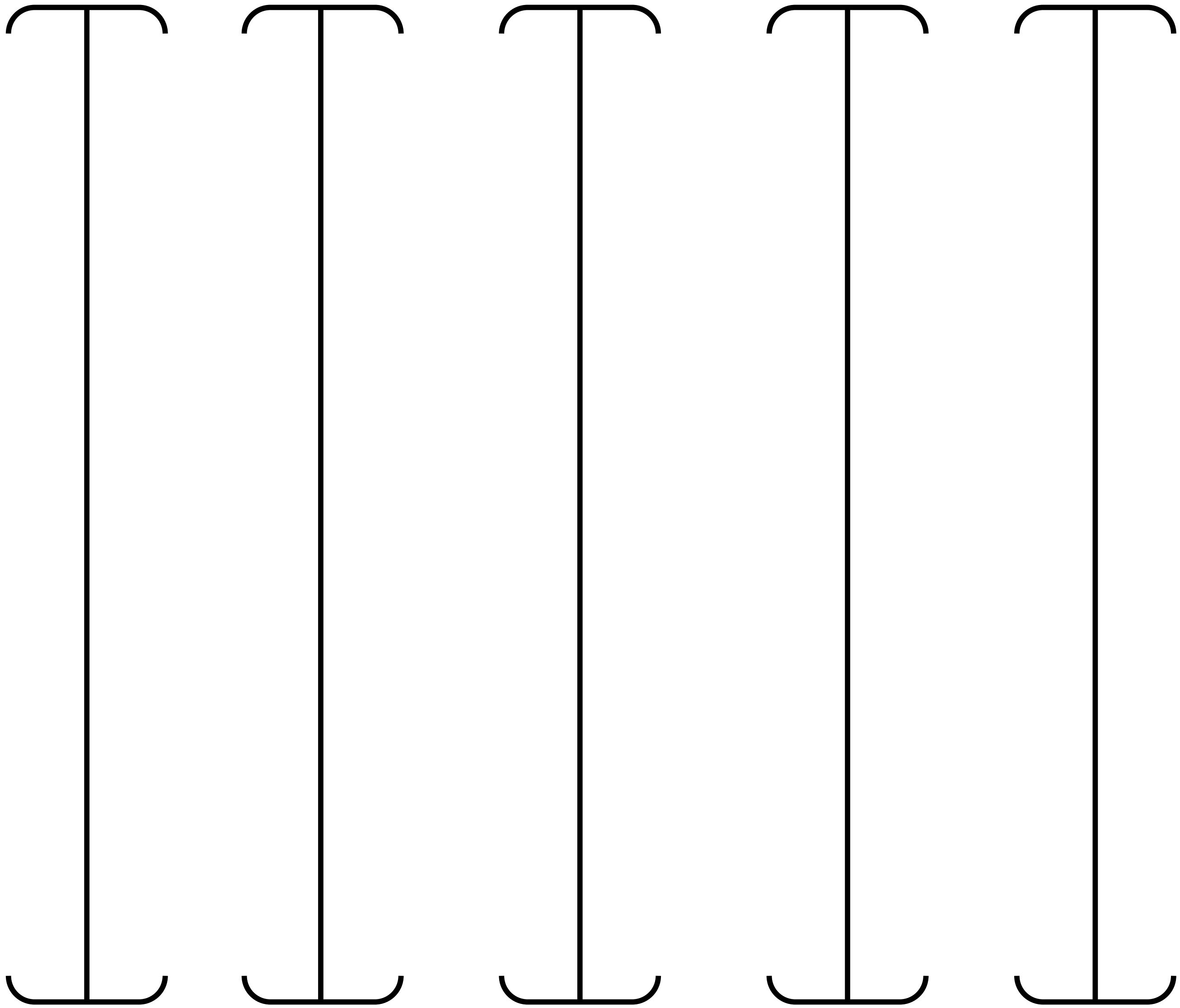
- Responsible for the demise of democracy?

The Upper Confidence Bound Algorithm

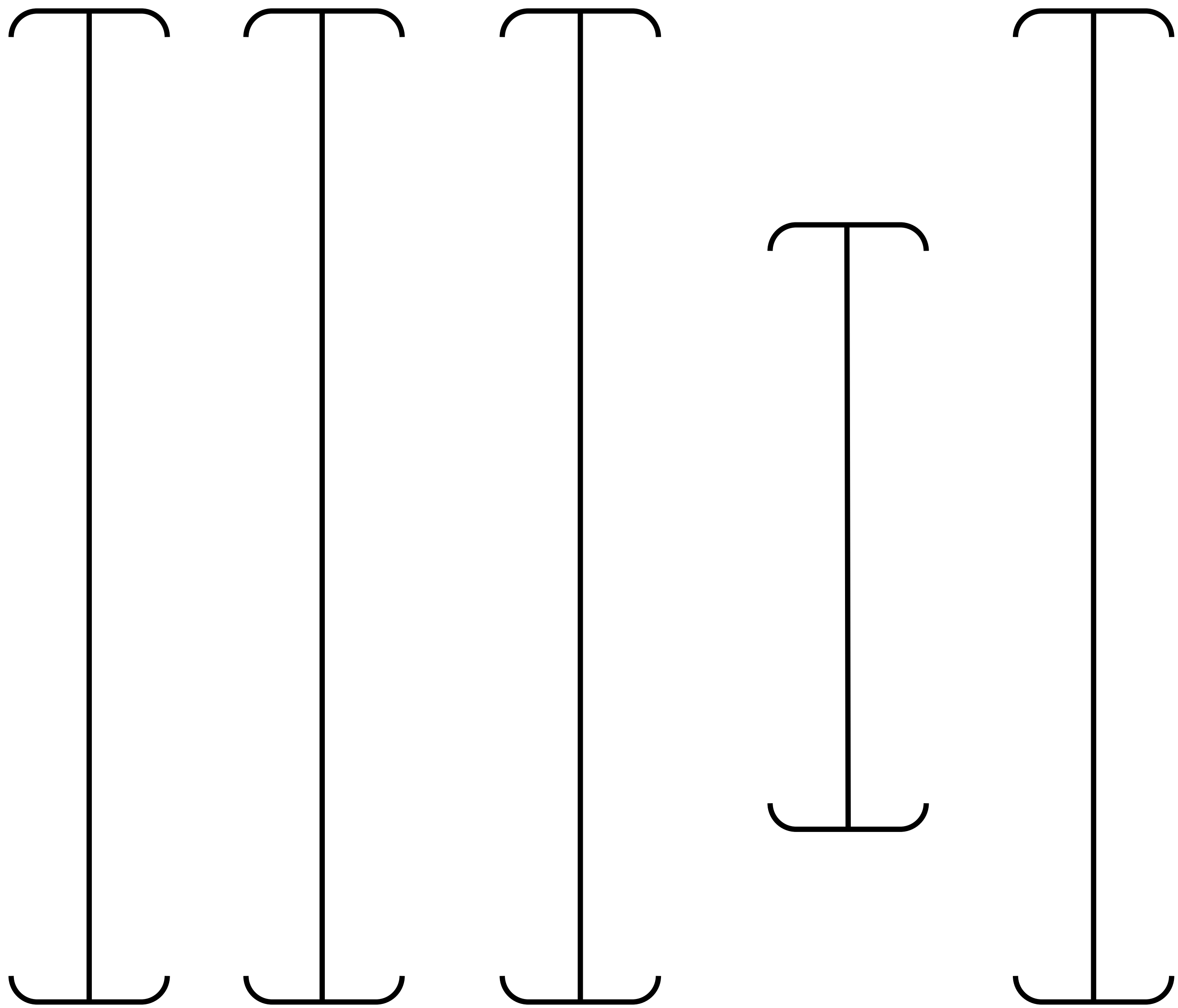
The Upper Confidence Bound Algorithm

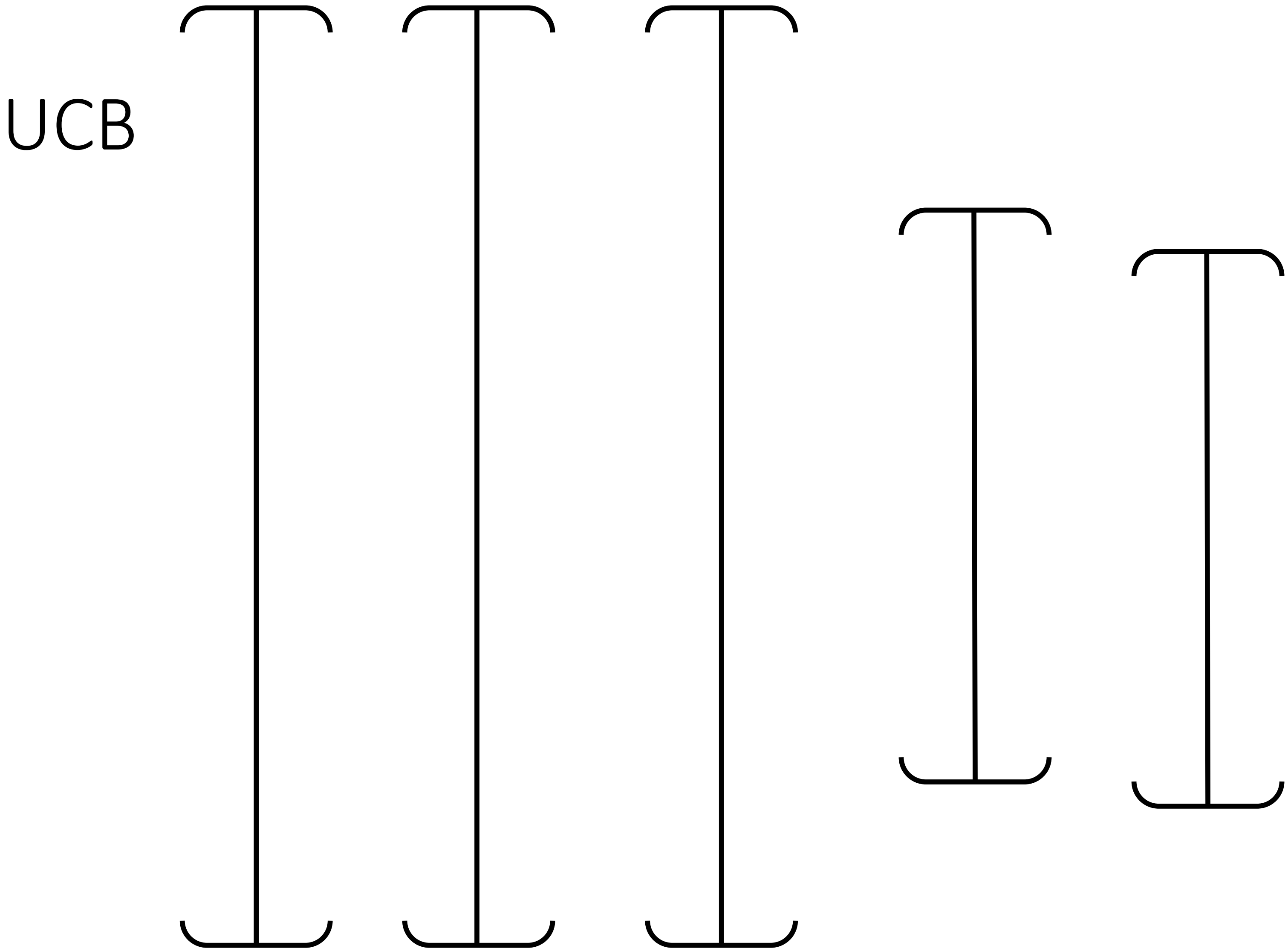
Starting phase – initialize all the arms

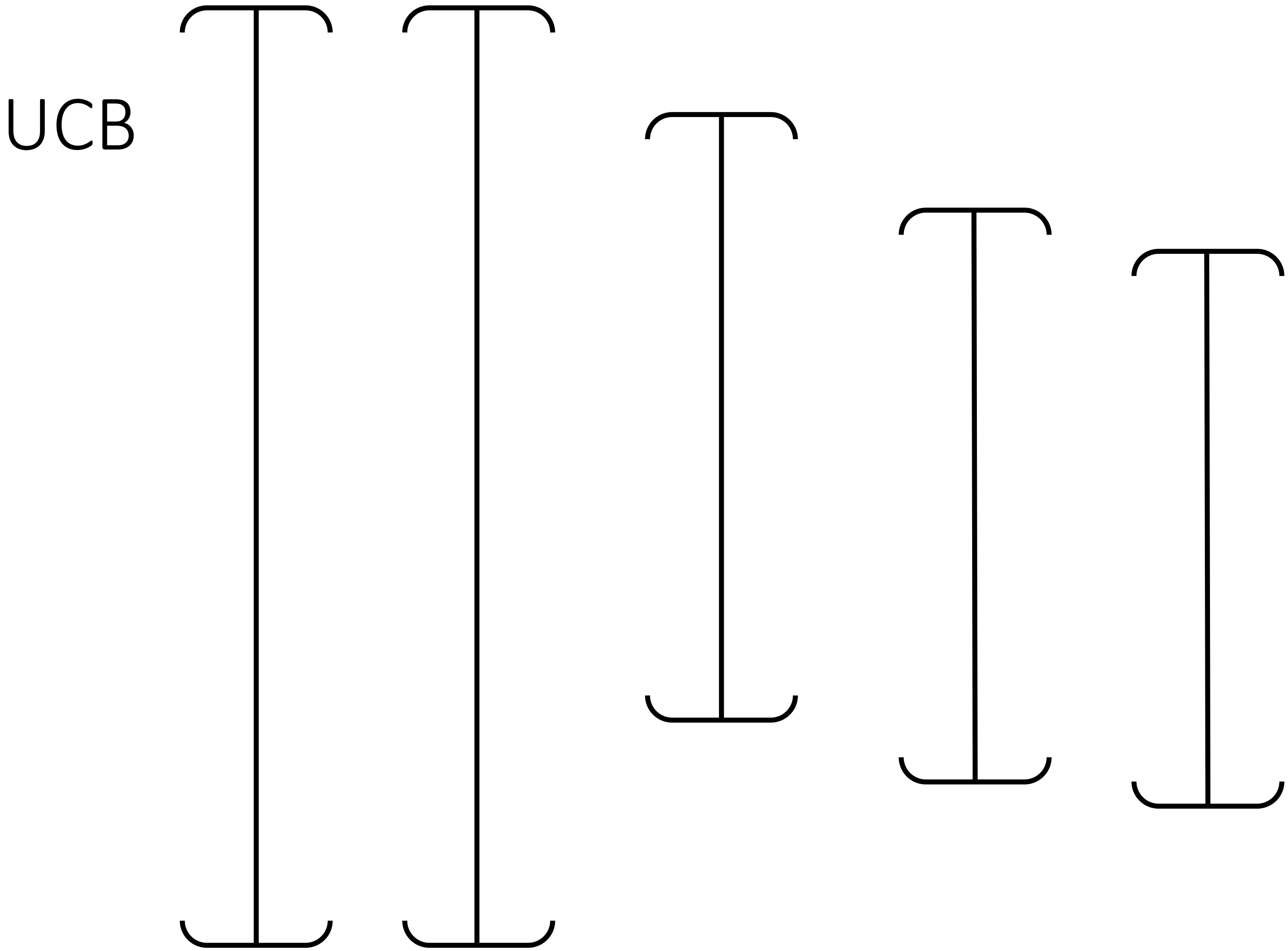
UCB

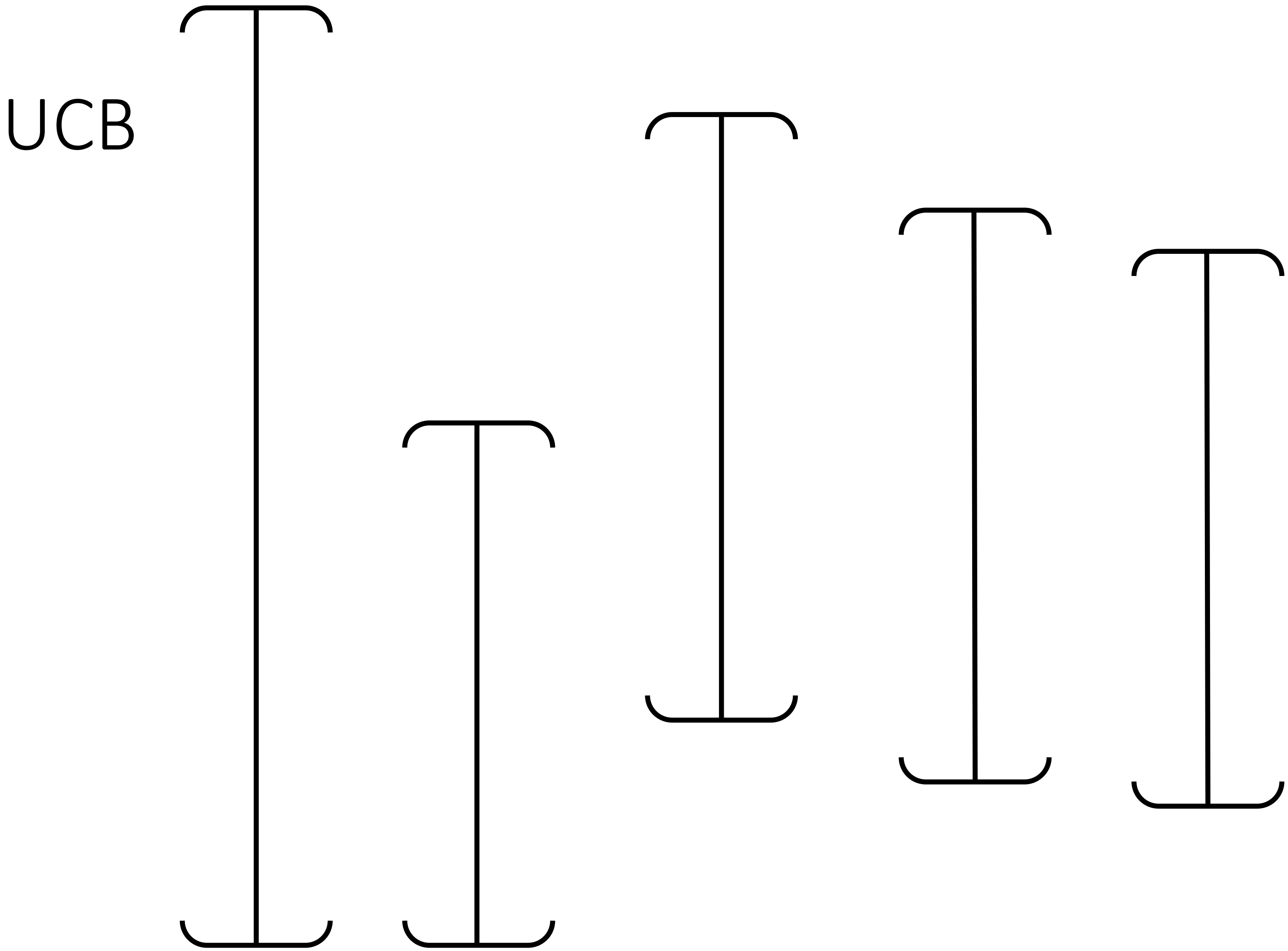


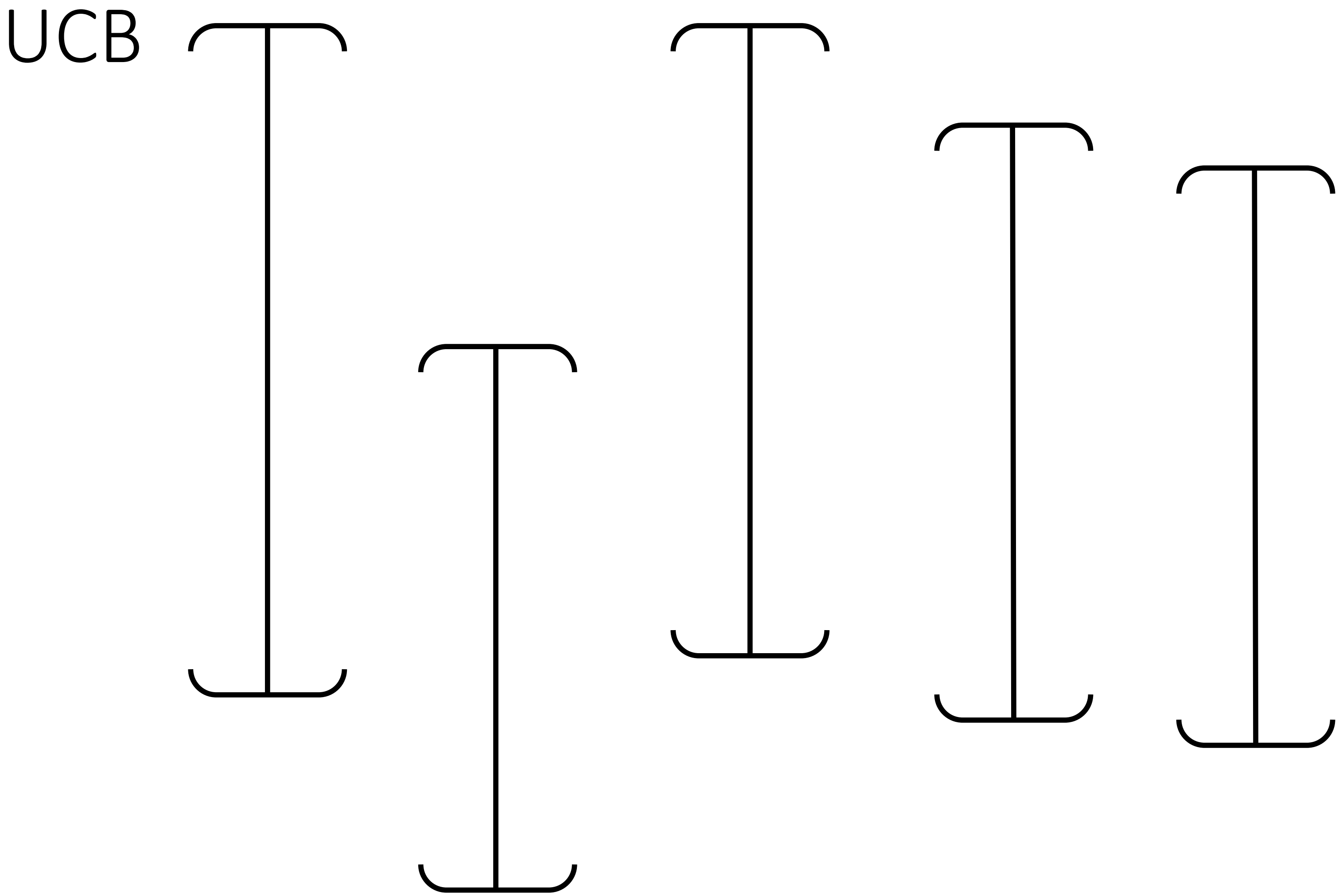
UCB




UCB


UCB

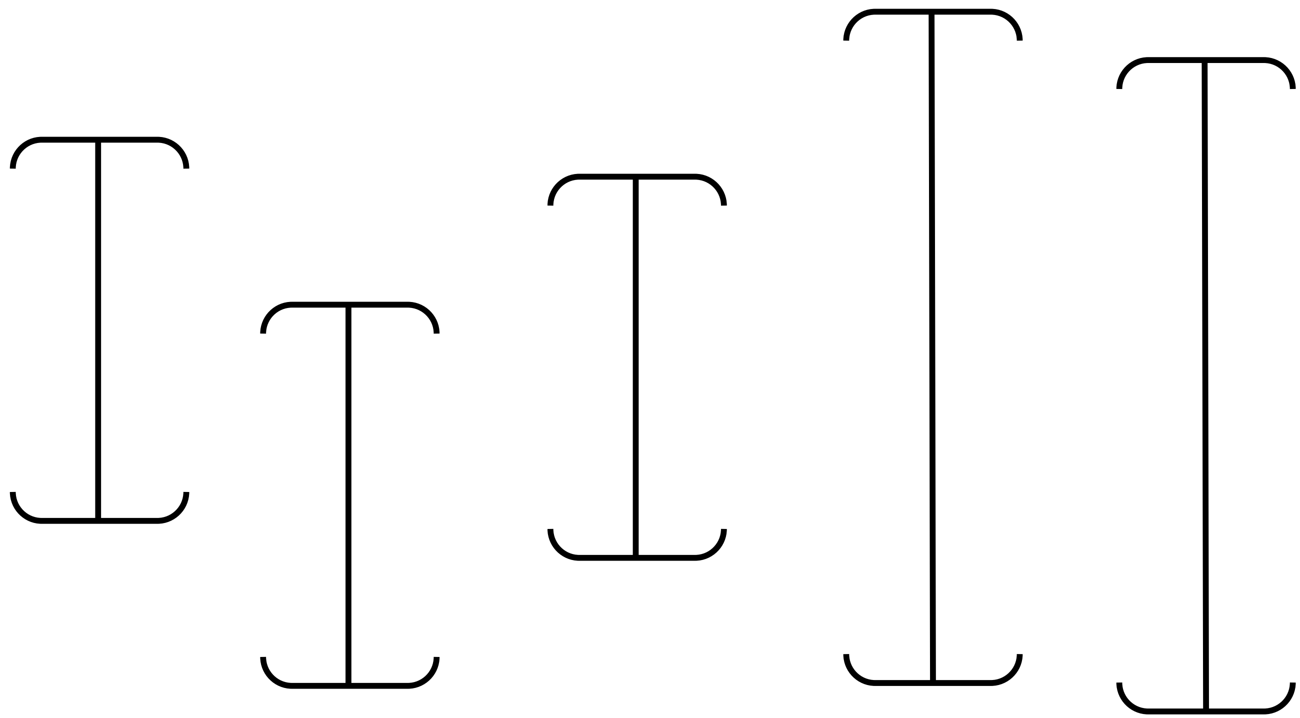

UCB


UCB

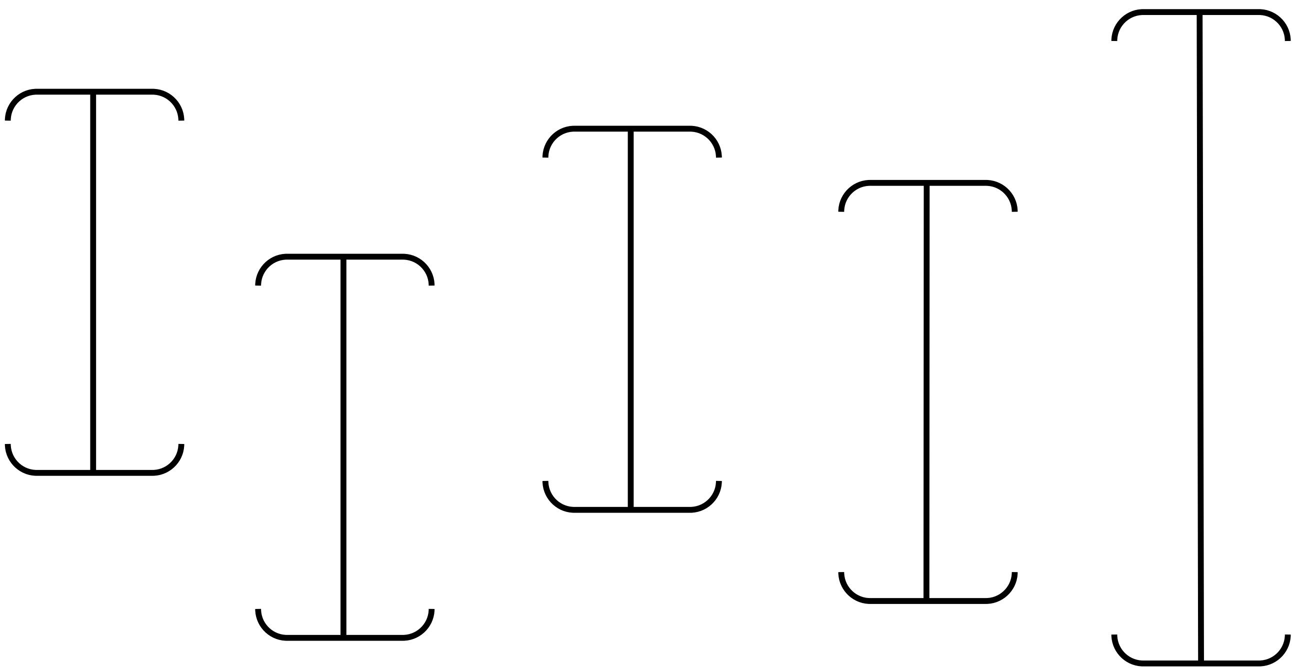
UCB

From now on: Choose the arm with the highest upper confidence bound

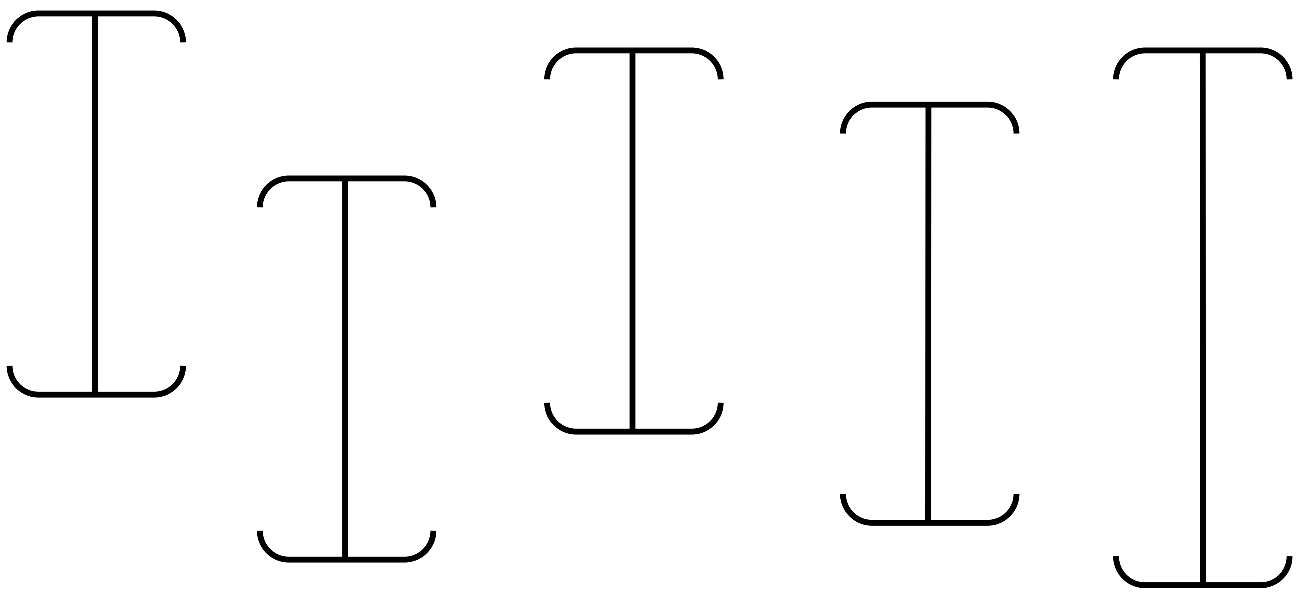
UCB



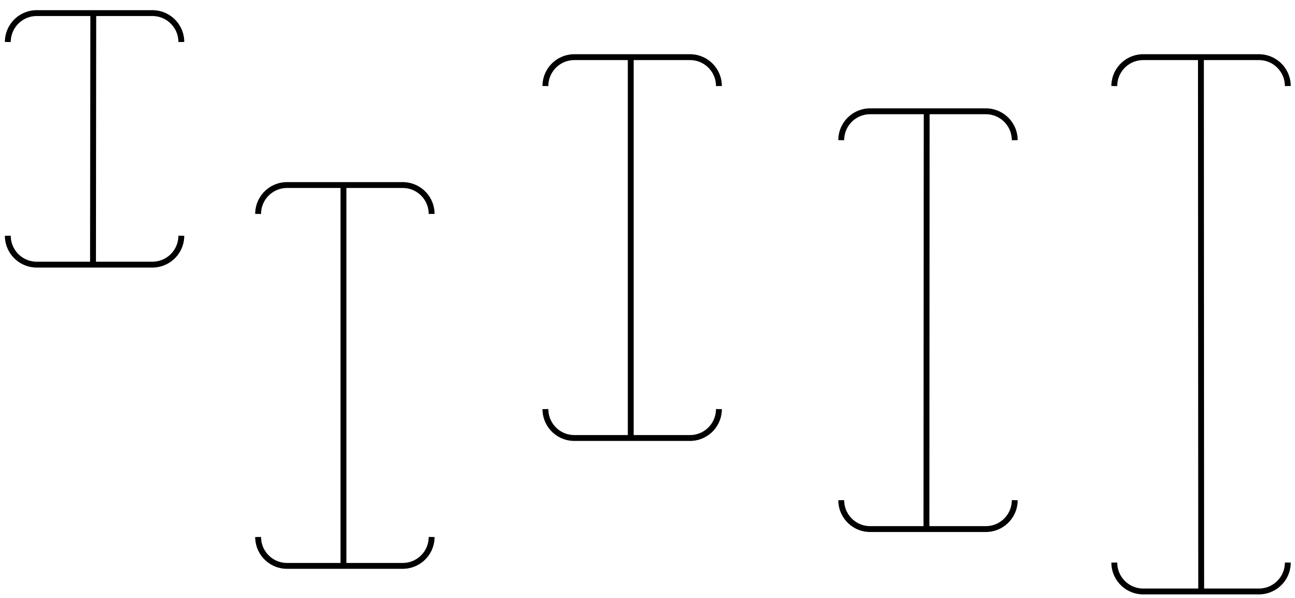
UCB



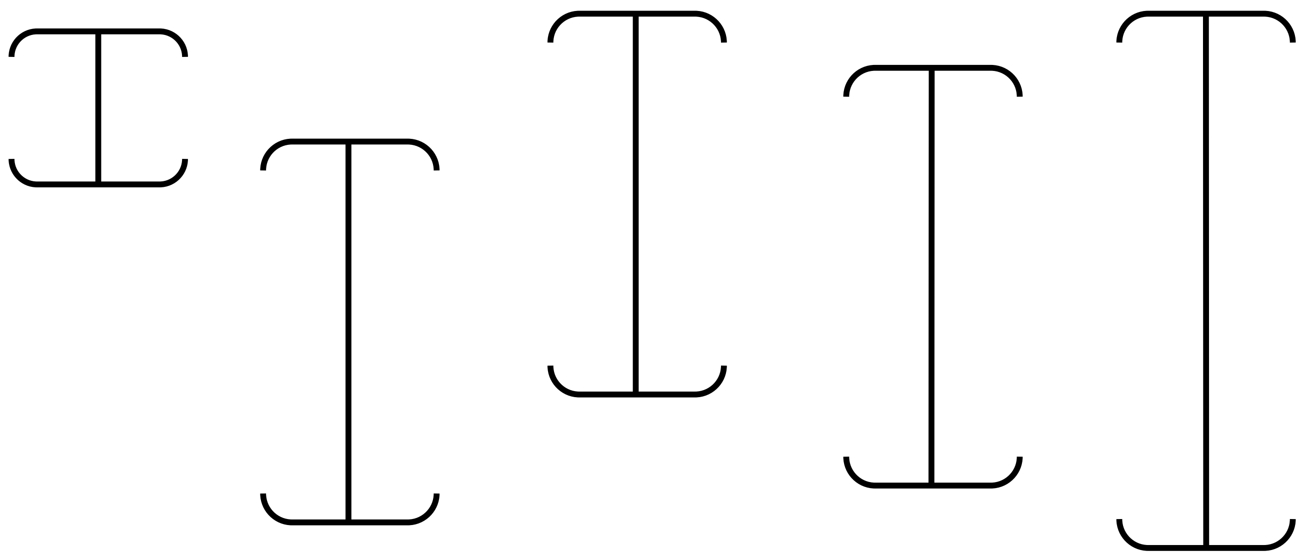
UCB



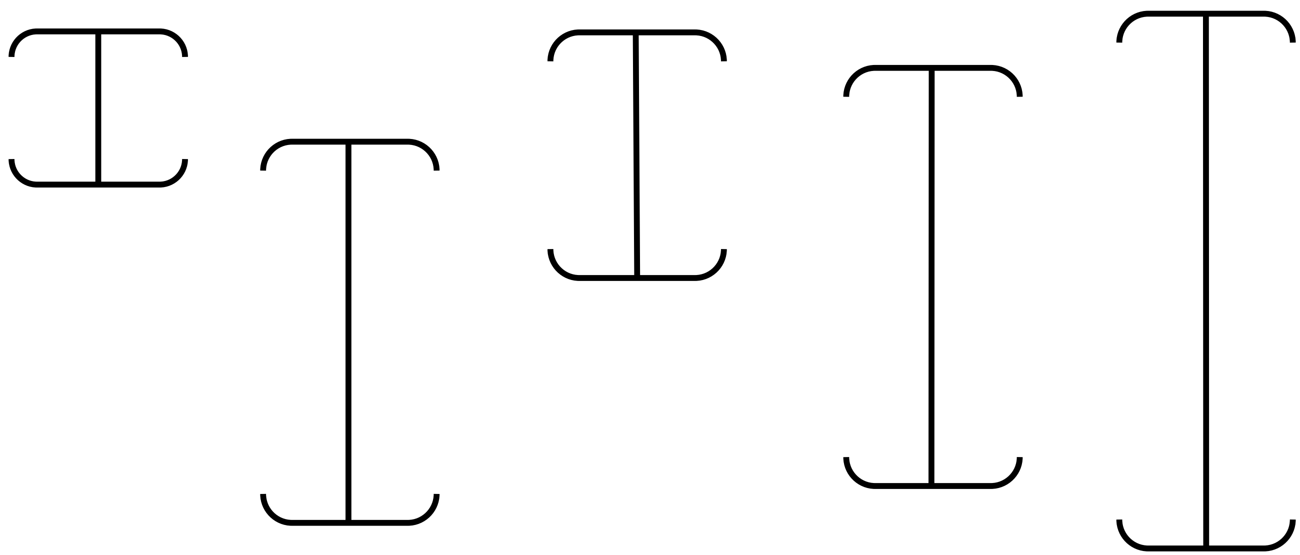
UCB



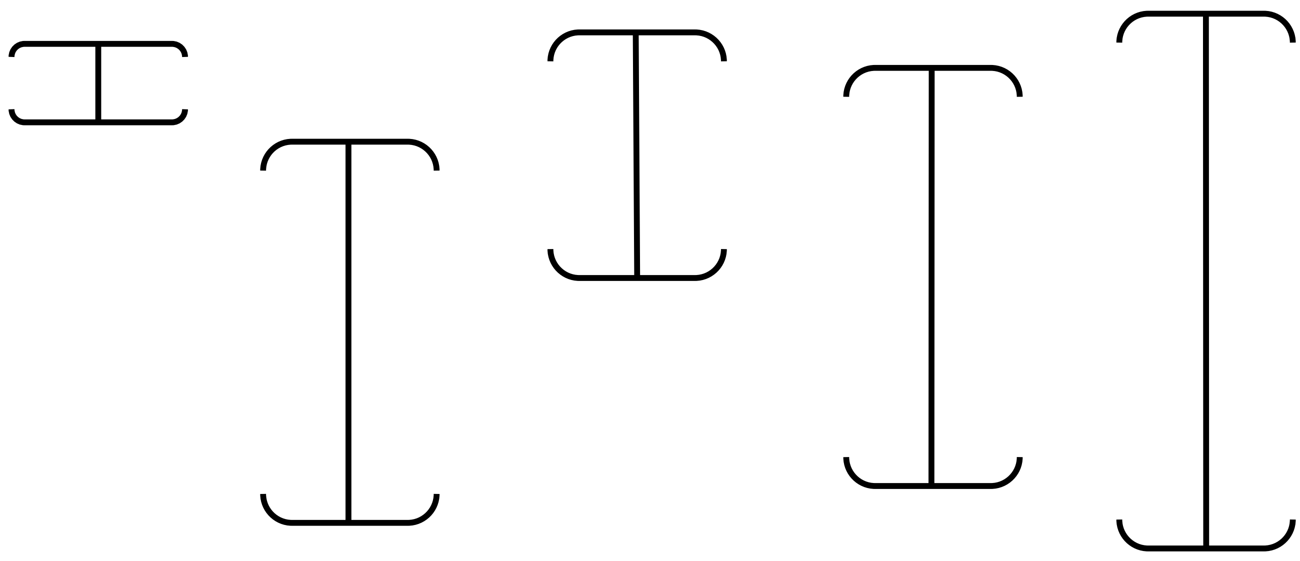
UCB



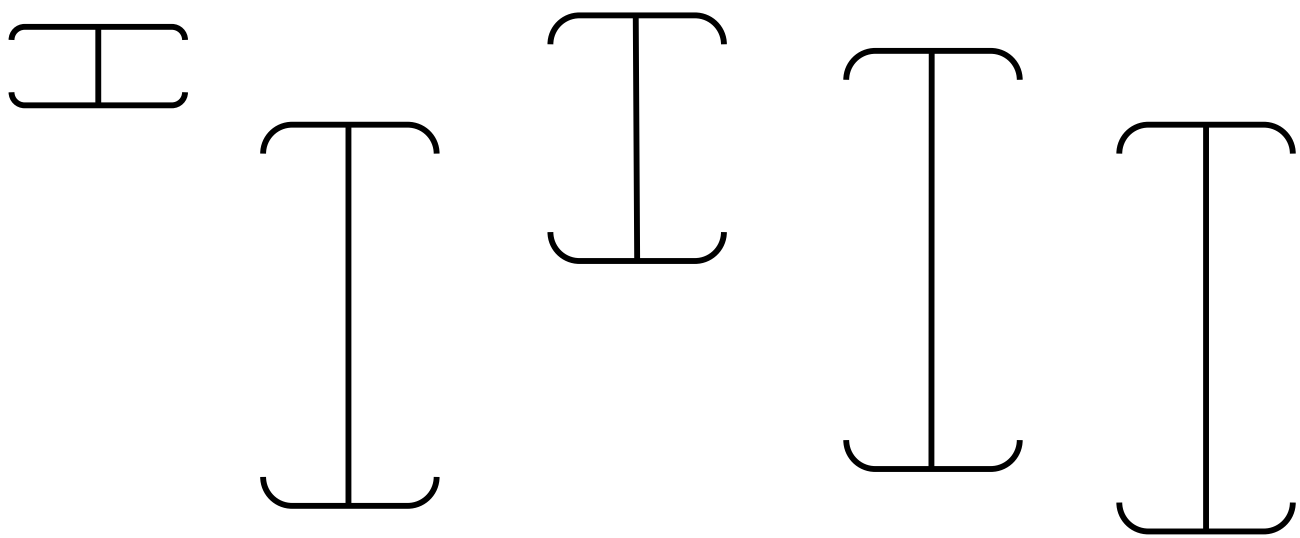
UCB



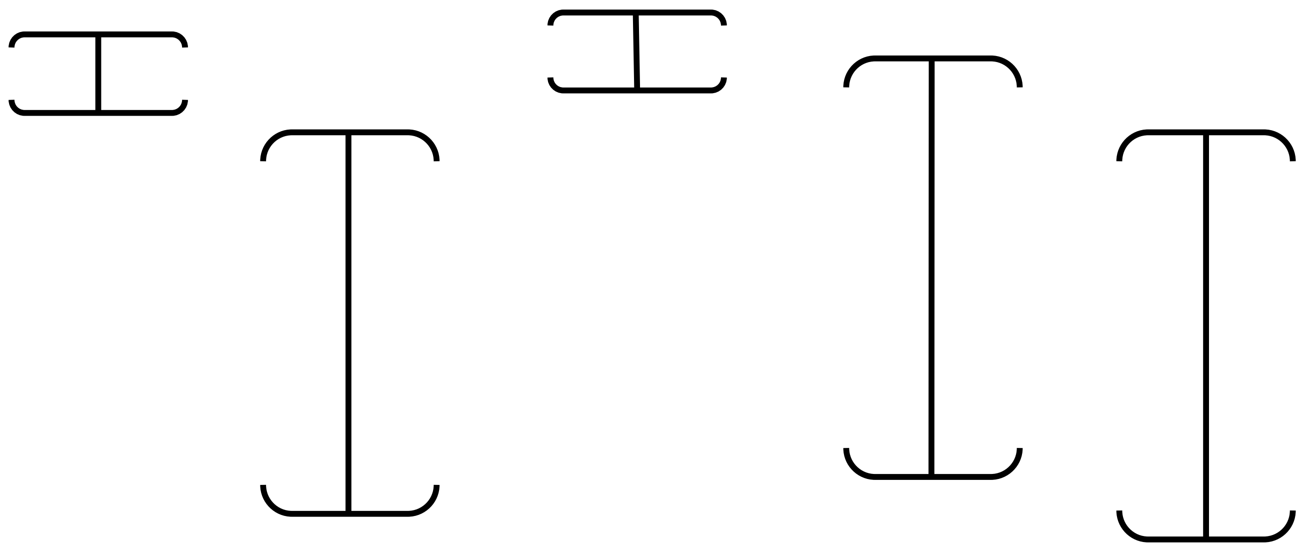
UCB



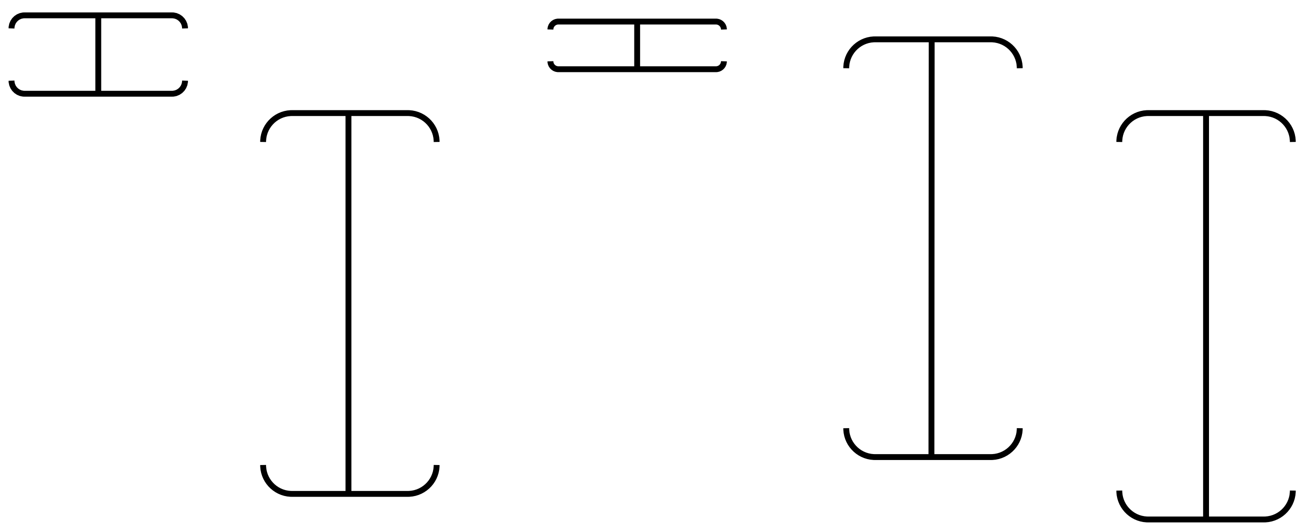
UCB



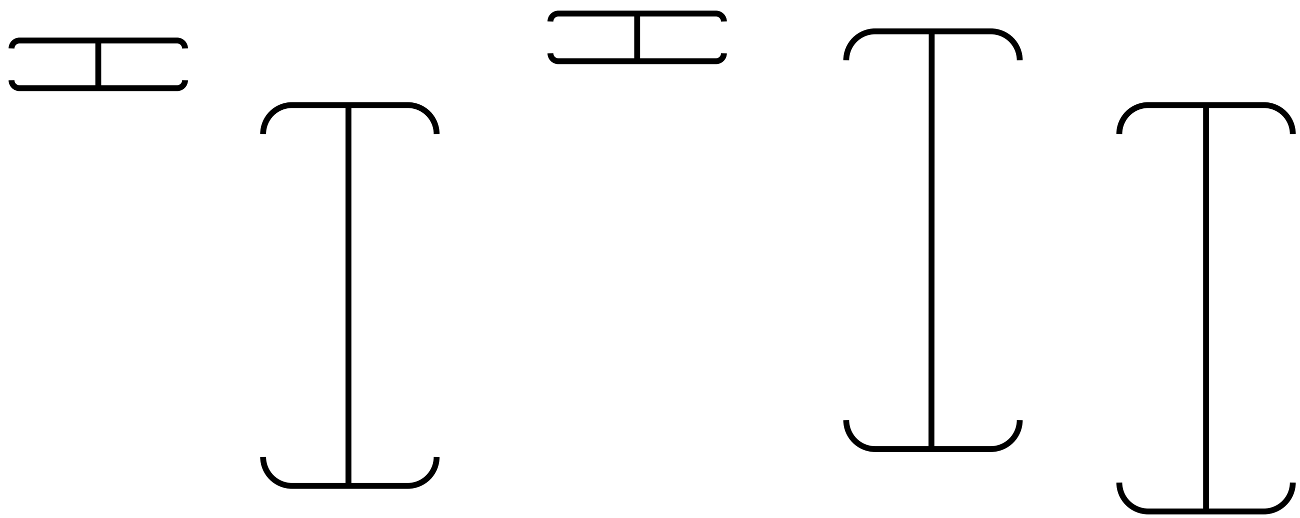
UCB



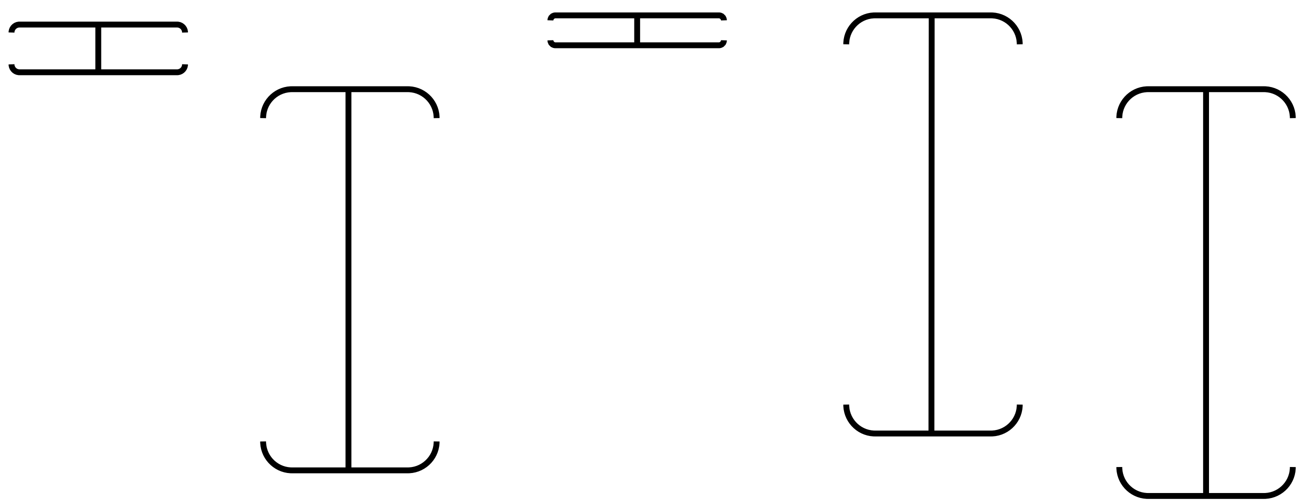
UCB



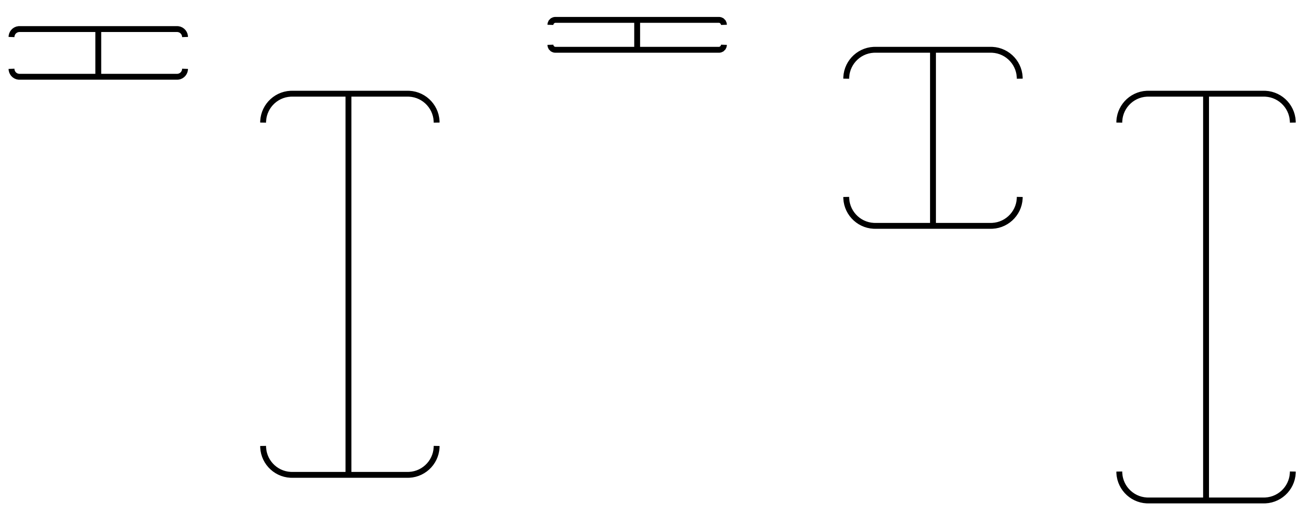
UCB



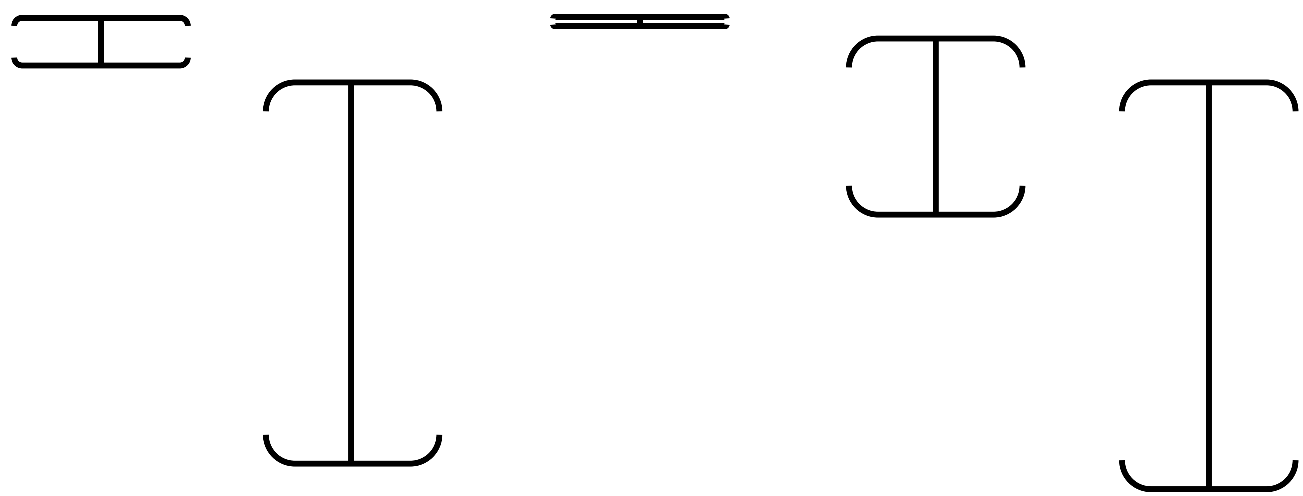
UCB



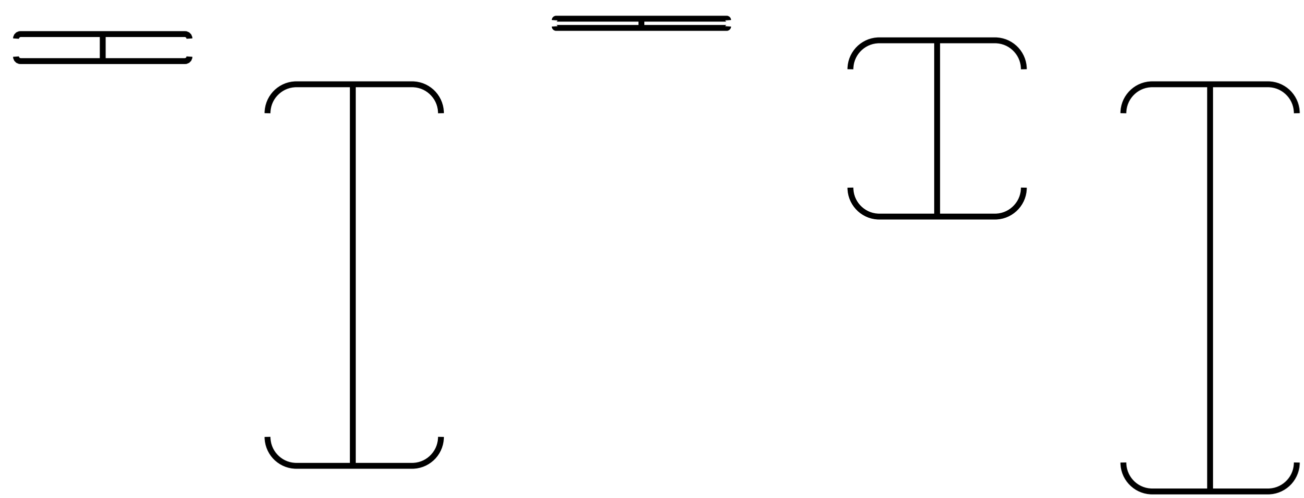
UCB



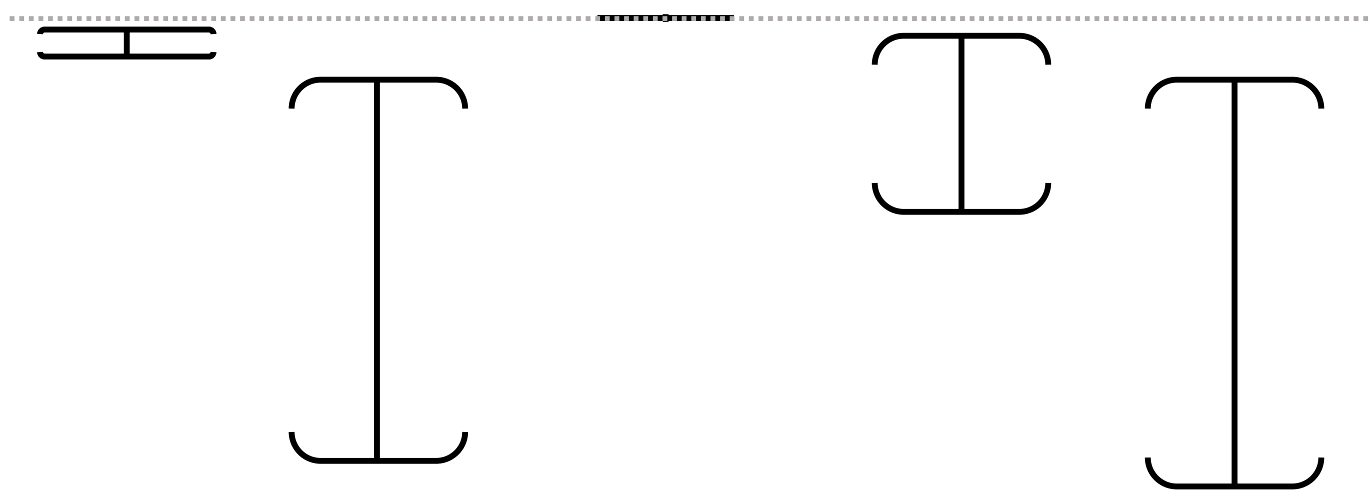
UCB



UCB



UCB



The ε-greedy algorithm

ε-greedy

Starting phase – initialize all the arms

ε-greedy

$\hat{X}_{t,2}$

$\hat{X}_{t,1}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

Randomly choose arms for a while (only exploration)

ε-greedy

$\hat{X}_{t,2}$

$\hat{X}_{t,1}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\widehat{X}_{t,1}$

$\widehat{X}_{t,2}$

$\widehat{X}_{t,3}$

$\widehat{X}_{t,5}$

$\widehat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$ $\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

With probability ε_t, play an arm uniformly at random.
(I’ll give you the formula for ε_t later. Think of it as constant/t.)

Otherwise, play the arm you think is the best.

ε-greedy

$\varepsilon_t = .85$

roll 1, explore

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$ $\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\varepsilon_t = .80$

roll 1, explore

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\varepsilon_t = .75$

roll 1, explore

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$ $\hat{X}_{t,2}$ $\hat{X}_{t,3}$ $\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\varepsilon_t = .70$

roll 0, exploit!

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$

$\hat{X}_{t,3}$

$\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$ $\hat{X}_{t,2}$ $\hat{X}_{t,3}$ $\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\varepsilon_t = .62$

roll 1, explore

$\hat{X}_{t,1}$ $\hat{X}_{t,2}$ $\hat{X}_{t,3}$ $\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\hat{X}_{t,1}$ $\hat{X}_{t,2}$ $\hat{X}_{t,3}$ $\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

After a while…

ε-greedy

$\varepsilon_t = .0001$
roll 0, exploit!

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$ $\hat{X}_{t,3}$ $\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy

$\varepsilon_t = .00001$

roll 0, exploit!

$\hat{X}_{t,1}$

$\hat{X}_{t,2}$ $\hat{X}_{t,3}$ $\hat{X}_{t,5}$

$\hat{X}_{t,4}$

ε-greedy formal statement

Input : number of rounds n, number of arms m, a constant k such that $k > \max\{10, \frac{4}{\min_j \Delta_j^2}\}$, sequence $\{\varepsilon_t\}_{t=1}^n = \min\left\{1, \frac{km}{t}\right\}$

Initialization: play all arms once and initialize $\widehat{X}_{j,t}$

for $t = m+1$ **to** n **do**

With probability ε_t play an arm uniformly at random (each arm has probability $\frac{1}{m}$ of being selected), otherwise (with probability $1-\varepsilon_t$) play ("best") arm j such that

$$\widehat{X}_{j,t-1} \geq \widehat{X}_{i,t-1} \ \forall i.$$

Get reward $X_j(t)$;

Update $\widehat{X}_{j,t}$;

end

UCB formal statement

Input : number of rounds n, number of arms m
Initialization: play all arms once and initialize $\widehat{X}_{j,t}$
for $t = m + 1$ **to** n **do**
  play arm j with the highest upper confidence bound on the mean estimate:

$$\widehat{X}_{j,t-1} + \sqrt{\frac{2\ \log(t)}{T_j(t-1)}};$$

Number of times arm *j* was played up to time *t*-1

  Get reward X_j;
  Update $\widehat{X}_{j,t}$;
end

Multi-armed bandit

Applications:

- Ad serving
  - Arms – possible ads
  - Reward – a click
- Website optimization
  - Arms – possible website options
  - Reward – user engagement
- Clinical Trials
  - Arms: possible medications
  - Reward: health outcomes

(Alternative to massive AB testing)

- Responsible for the demise of democracy?

Multi-armed Bandits
Part 2: Theory

Cynthia Rudin

Duke University

$$\widehat{X}_{j,t} = \frac{1}{T_j(t-1)} \sum_{s=1}^{t-1} X_j(s) \overbrace{\mathbb{1}_{\{I_t = j\}}}$$ = Estimate of mean reward for our strategy I_t

Number of times arm j was played up to time t-1

μ_j = expected reward for arm j

$$R_n^{(\text{raw})} = \sum_{t=1}^{n} \sum_{j=1}^{m} [X^*(t) - X_j(t)] \mathbb{1}_{\{I_t = j\}}$$ = raw regret for playing our strategy I instead of always playing the best arm "*".

$$R_n = \sum_{t=1}^{n} \sum_{j=1}^{m} \underbrace{[\mu_* - \mu_j]} \mathbb{1}_{\{I_t = j\}}$$ = regret for playing our strategy, using arms' mean rewards.

$$R_n = \sum_{t=1}^{n} \sum_{j=1}^{m} \overbrace{\Delta_j} \mathbb{1}_{\{I_t = j\}}$$

expected regret for playing our strategy

$$\mathbb{E}[R_n] = \mathbb{E}\sum_{t=1}^{n}\sum_{j=1}^{m}\Delta_j \mathbb{1}_{\{I_t=j\}} = \sum_{j=1}^{m}\Delta_j \underbrace{\mathbb{E}\left[T_j(n)\right]}$$

We want to bound the expected regret of our strategy

$$R_n = \sum_{t=1}^{n}\sum_{j=1}^{m}\underbrace{[\mu_* - \mu_j]}\mathbb{1}_{\{I_t=j\}}$$

= regret for playing our strategy, using arms' mean rewards.

$$R_n = \sum_{t=1}^{n}\sum_{j=1}^{m}\overbrace{\Delta_j}\mathbb{1}_{\{I_t=j\}}$$

ε-greedy formal statement

Input : number of rounds n, number of arms m, a constant k such that $k > \max\{10, \frac{4}{\min_j \Delta_j^2}\}$ sequence $\{\varepsilon_t\}_{t=1}^n = \min\left\{1, \frac{km}{t}\right\}$ $\widehat{X}_{j,t}$

Initialization: play all arms once and initialize $\widehat{X}_{j,m}$ (defined in (1)) for each $j = 1, \cdots, m$

for $t = m + 1$ **to** n **do**

With probability ε_t play an arm uniformly at random (each arm has probability $\frac{1}{m}$ of being selected), otherwise (with probability $1 - \varepsilon_t$) play ("best") arm j such that

$$\widehat{X}_{j,t-1} \geq \widehat{X}_{i,t-1} \ \forall i.$$

Get reward $X_j(t)$;

Update $\widehat{X}_{j,t}$;

end

Regret bound for ε-greedy

probability to explore

probability to choose j when exploring

Regret bound for ε-greedy

Expected regret

Starting phase regret bound

Regret for arm j

Prob to exploit

$$\mathbb{E}[R_n] \leq ekm^2 + \sum_{t=e^2km+1}^{n} \sum_{j:\mu_j<\mu_*} \Delta_j \left(\varepsilon_t \frac{1}{m} + (1-\varepsilon_t)\beta_j(t)\right)$$

where

bound on Prob you think j is the best when it's not!

$$\beta_j(t) = k\left(\frac{t}{mke}\right)^{-\frac{k}{10}} \log\left(\frac{t}{mke}\right) + \frac{4e^{\frac{1}{2}}}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}.$$

$$\{\varepsilon_t\}_{t=1}^{n} = \min\left\{1, \frac{km}{t}\right\}$$

$$k > \max\{10, \frac{4}{\min_j \Delta_j^2}\}$$

Regret bound for ε-greedy

Logarithmic in number of rounds, n

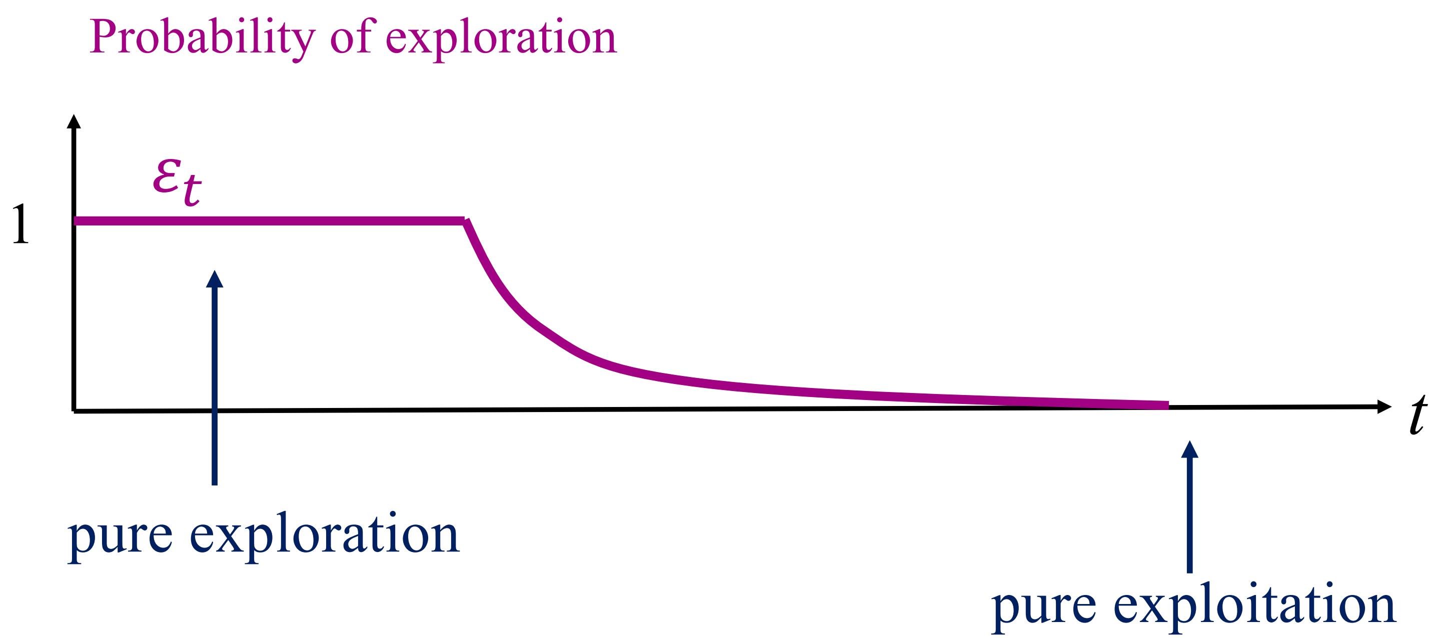
Expected regret

$$\mathbb{E}[R_n] \leq ekm^2 + \sum_{t=e^2km+1}^{n} \sum_{j:\mu_j<\mu_*} \Delta_j \left(\varepsilon_t \frac{1}{m} + (1-\varepsilon_t)\beta_j(t)\right)$$

where

$$\beta_j(t) = k\left(\frac{t}{mke}\right)^{-\frac{k}{10}} \log\left(\frac{t}{mke}\right) + \frac{4e^{\frac{1}{2}}}{\Delta_j^2}\left(\frac{t}{mke}\right)^{-\frac{k\Delta_j^2}{4}}.$$

$$\{\varepsilon_t\}_{t=1}^{n} = \min\left\{1, \frac{km}{t}\right\}$$



UCB formal statement

Input : number of rounds n, number of arms m

Initialization: play all arms once and initialize $\widehat{X}_{j,t}$

No parameters!

for $t = m + 1$ **to** n **do**

play arm j with the highest upper confidence bound on the mean estimate:

$$\widehat{X}_{j,t-1} + \sqrt{\frac{2\log(t)}{T_j(t-1)}};$$

Number of times arm j was played up to time t-1

Get reward X_j;

Update $\widehat{X}_{j,t}$;

end

Regret bound for UCB

Regret bound for UCB

$$\mathbb{E}[R_n] \leq \sum_{j=1}^{m} \Delta_j + \sum_{j:\mu_j<\mu_*} \frac{8}{\Delta_j} \log(n) + \sum_{j=1}^{m} \Delta_j \left(1 + \sum_{t=m+1}^{n} 2t^{-4}(t-1-m)^2\right)$$

Logarithmic in number of rounds, *n*

Notes

- Both algorithms have regret that increases only logarithmically in the number of rounds. Proofs are in the notes.
- There are theorems that do not involve Δ_j's. (One is in the notes.)
- Both algorithms are about equally good in practice.

Multi-armed Bandits
Part 3: Contextual Bandits

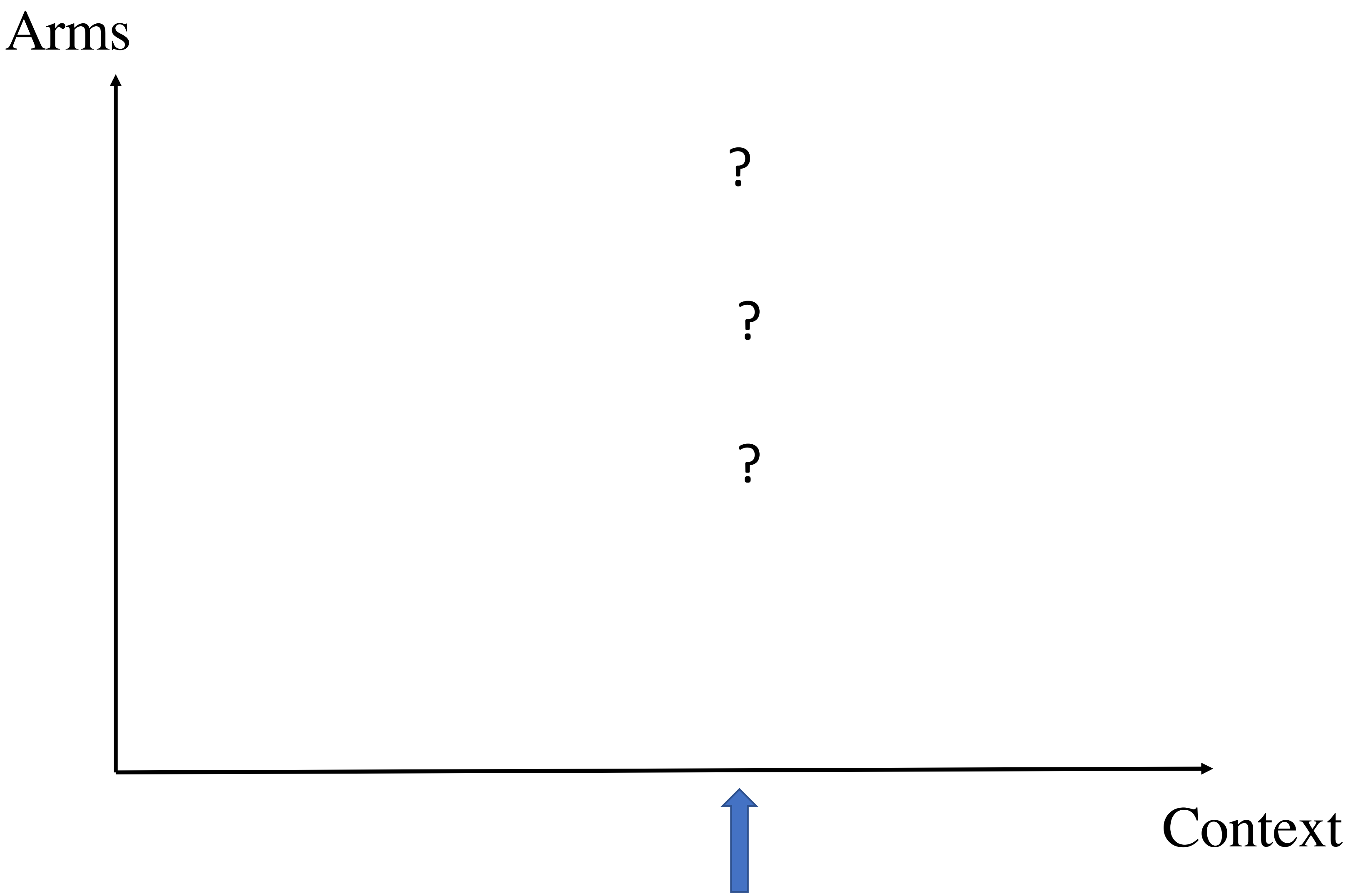
Cynthia Rudin

Duke University

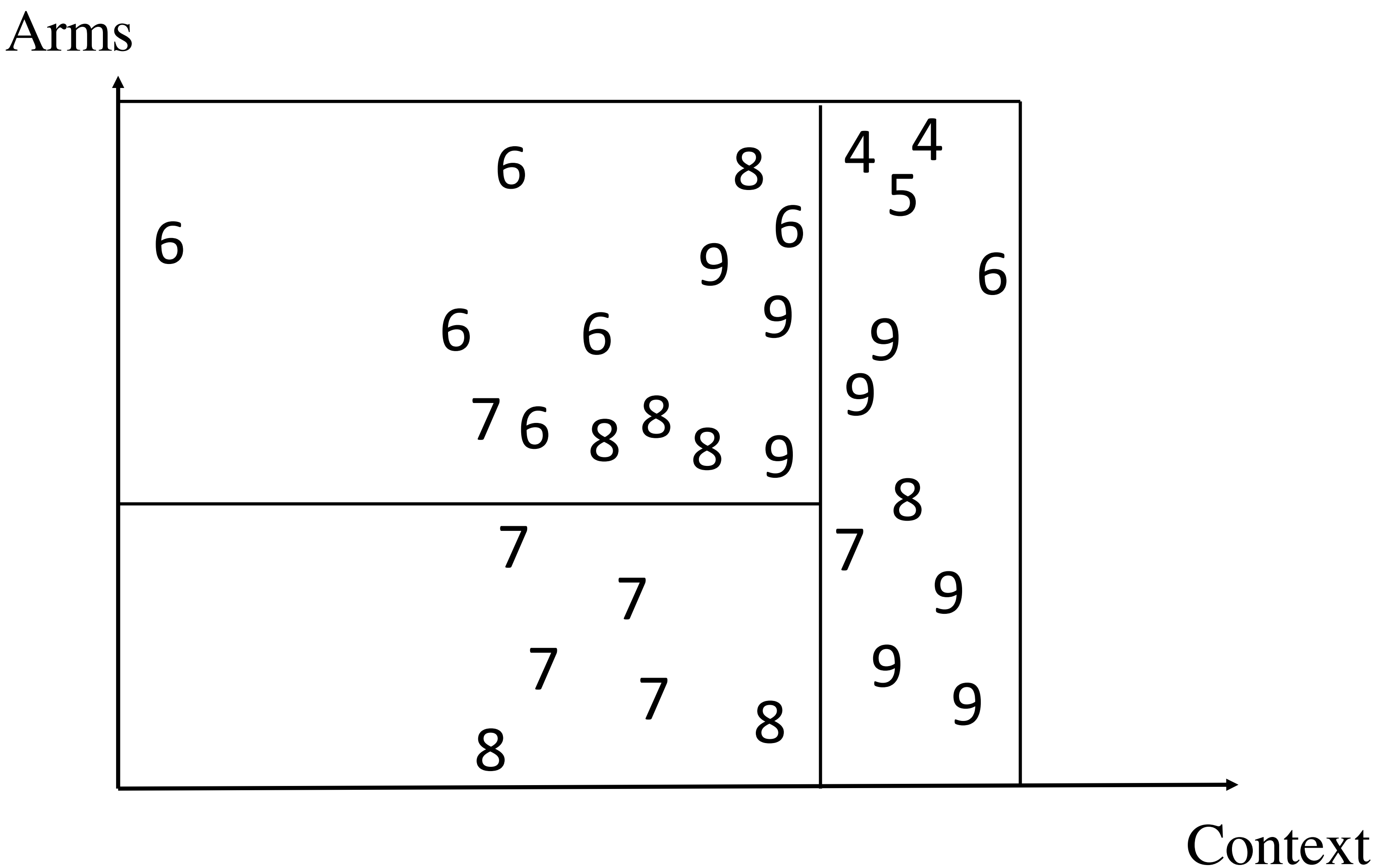
Context

$$\text{user\_in\_context} = \begin{pmatrix} \text{age} \\ \text{number of FaceBook friends} \\ \text{estimated IQ} \\ 1_{\text{if introvert}} \\ 1_{\text{if likes jazz}} \\ 1_{\text{if it is between 12am and 6am}} \\ 1_{\text{if browsing dating sites}} \end{pmatrix}$$

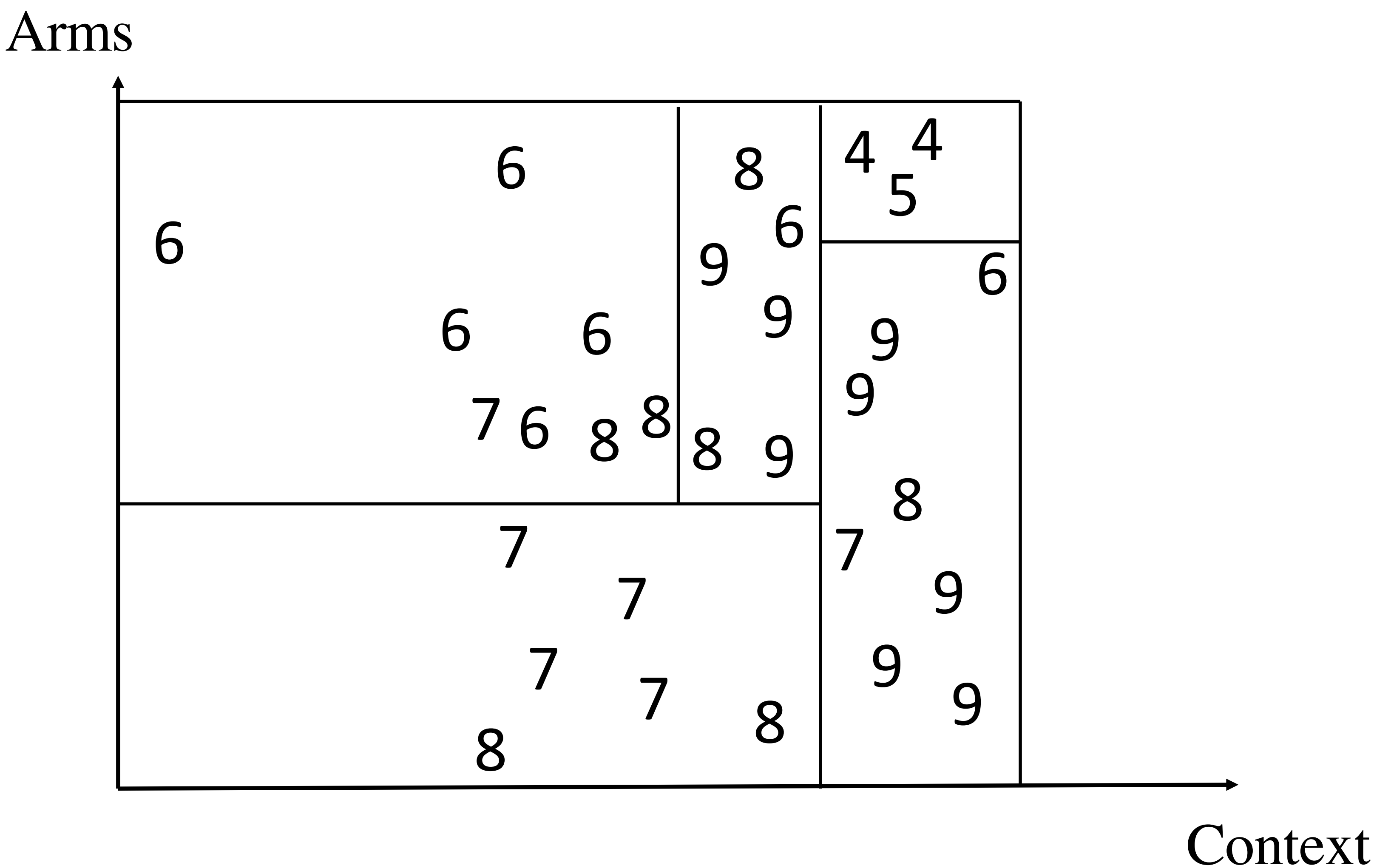
arms

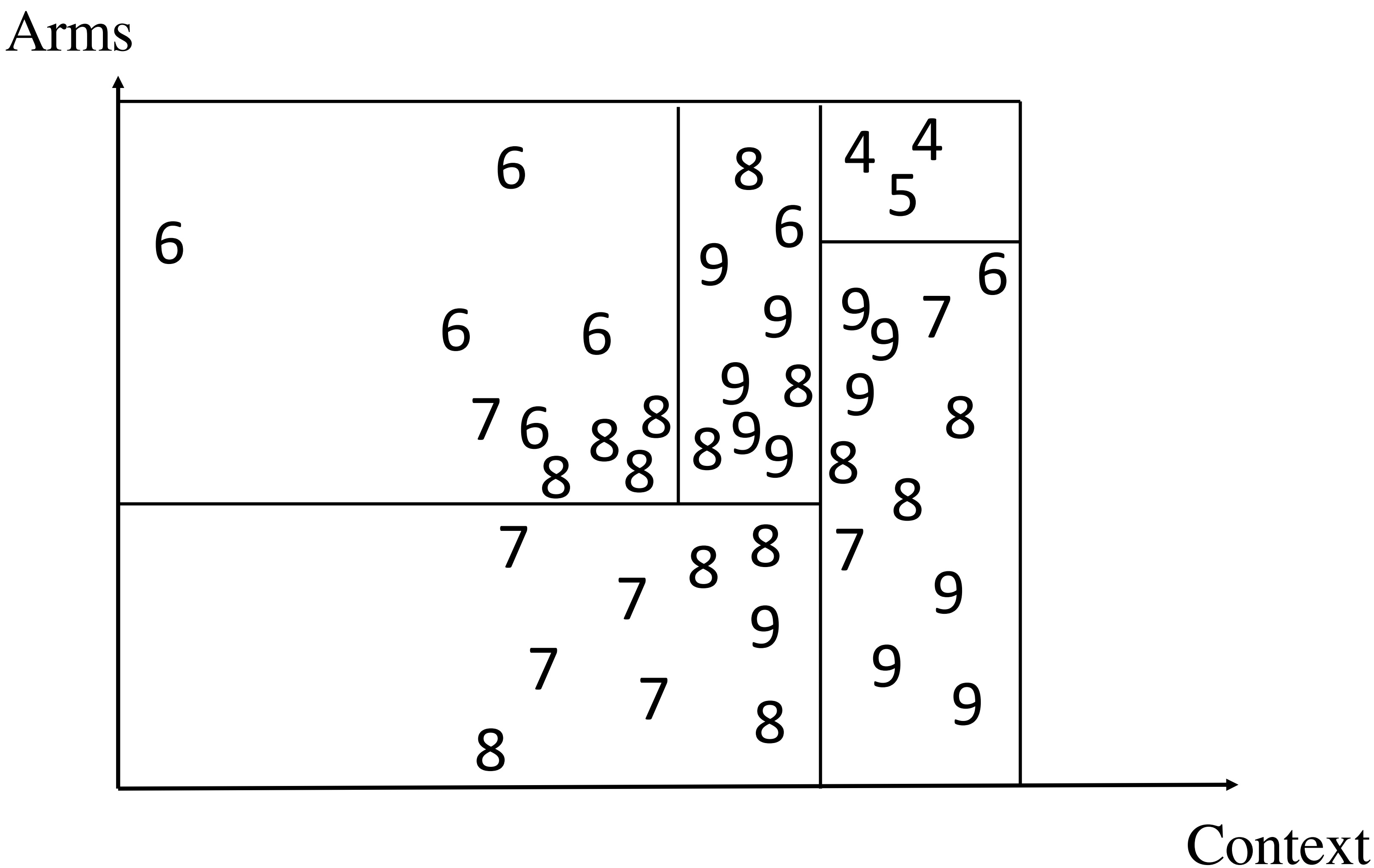


Arms
?
?
?
Context

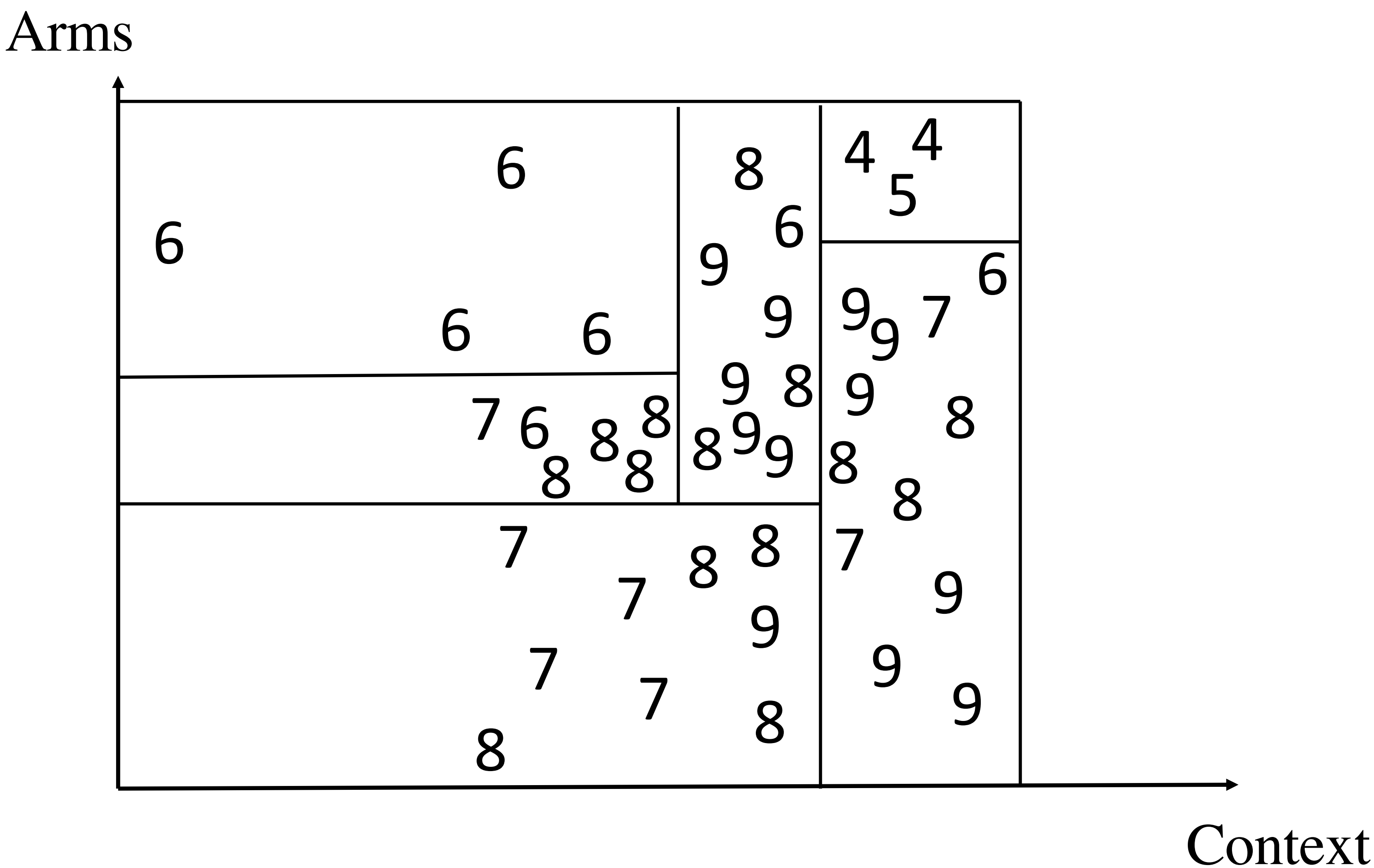


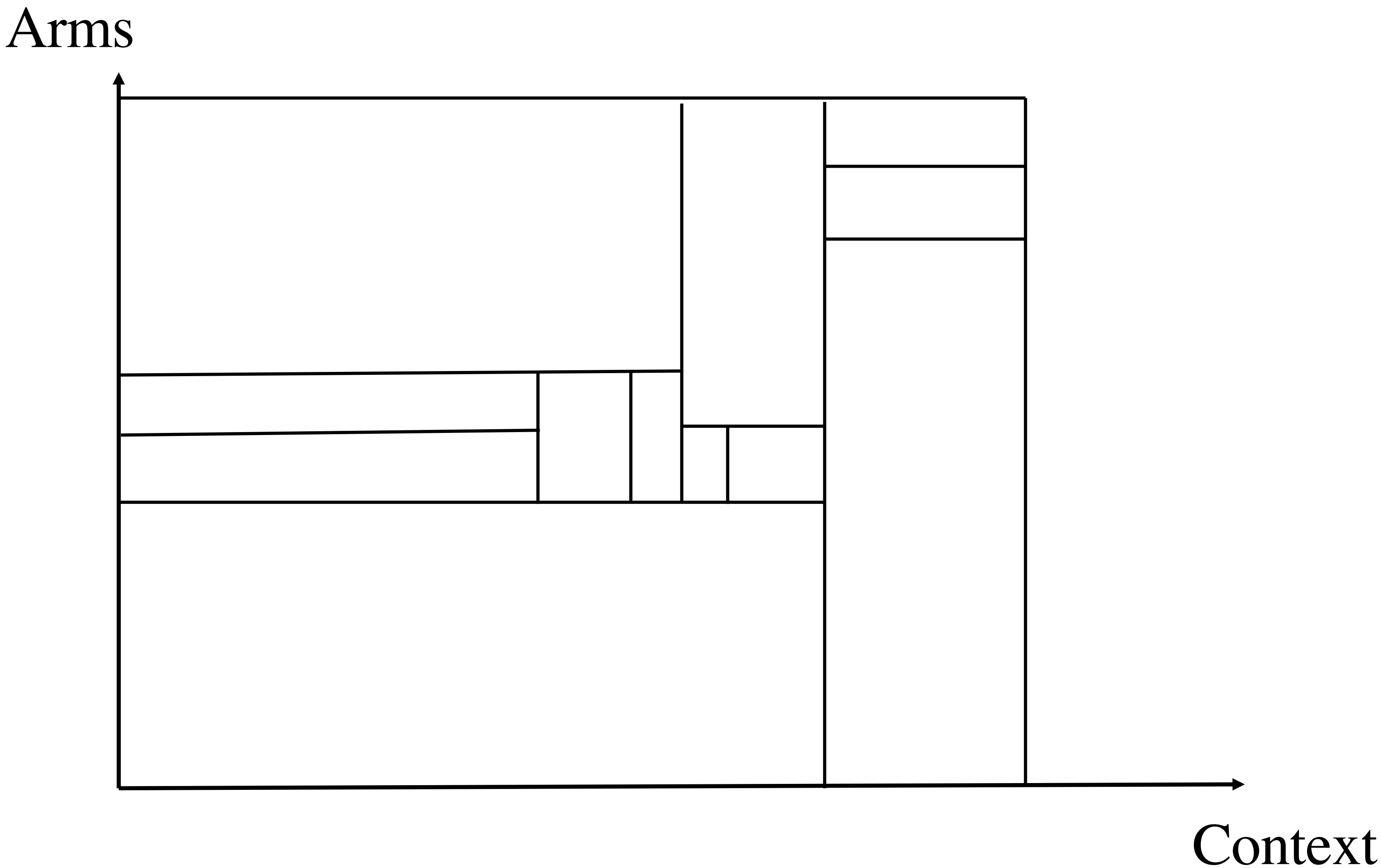
Arms
6
6
8
4
4
5
6
9
6
6
6
9
9
9
7
6
8
8
8
9
8
7
7
7
9
7
7
9
8
9
8
Context


Arms
6
6
6
6
7 6 8 8
8
6
9
9
8 9
4 4
5
6
9
9
8
7
9
9
9
7
7
7
7
8
8
Context



Arms
Context


Arms
Context


Arms
Context

How The New York Times is Experimenting with Recommendation Algorithms

NYT Open

Algorithmic curation at The Times is used in designated parts of our website and apps.

Anna Coenen Follow
Oct 17, 2019 · 6 min read

A contextual recommendation approach

One recommendation approach we have taken uses a class of algorithms called contextual multi-armed bandits. Contextual bandits learn over time how people engage with particular articles. They then recommend articles that they predict will garner higher engagement from readers. The *contextual* part means that these bandits can use additional information to get a better estimate of how engaging an article might be to a particular reader. For example, they can take into account a reader's geographical region (like country or state) or reading history to decide if a particular article would be relevant to that reader.

[“recommended”: “article B”; “reader state”: “Texas”, “clicked”: “yes”]
[“recommended”: “article A”; “reader state”: “New York”, “clicked”: “yes”]
[“recommended”: “article B”; “reader state”: “New York”, “clicked”: “no”]
[“recommended”: “article B”; “reader state”: “California”, “clicked”: “no”]
[“recommended”: “article A”; “reader state”: “New York”, “clicked”: “no”]

Once the bandit has been trained on the initial data, it might suggest Article A, Article B or a new article, C, for a new reader from New York. The bandit would be most likely to recommend Article A because the article had the highest click-through rate with New York readers in the past. With some smaller probability, it might also try showing Article C, because it doesn't yet know how engaging it is and needs to generate some data to learn about it.

Lots of bandits

- Sleeping bandits
- Mortal bandits
- Bandits where the mean rewards are nonstationary
- Bandits with arms that lock for a while
- Bandits with delayed rewards

: