INTRO TO DECISION-MAKING: BANDITS 3

NOTATION

REVIEW

X — Capital letters (usually) a random variable

$\mathscr{X}$ — Calligraphic letters are sets

$x \in \mathscr{X}$, q_* — lowercase letters are elements of a set or functions

$$\sum_{x \in \mathscr{X}} \Pr(X = x)$$

NOTATIOIN

REVIEW

A_t — Random variable for action that happens at time t

a_t — Instantiation of A_t, i.e., the outcome has been observed

R_t — Random variable for a reward at a time t. It depends on A_t.

R_n — Random variable for reward observed for a particular arm.

- The reward for each arm is a different random variable.
- $R_n^{(i)}$ — Random variable for the reward for the n^{th} time the i^{th} arm is tried. (Implicit)

Q_n — Random variable for value estimate of a particular arm

PROBABILITIES

ACTION SELECTION

$\epsilon-$Greedy action selection $\Pr(A_t = a)$

$X \in \{\text{greedy}, \text{random}\}$ — Random variable representing

If $X = \text{random}$ # $\Pr(X = \text{random}) = \epsilon$

$A_t \sim U(\mathscr{A})$ # $\Pr(A_t = a \,|\, X = \text{random}) = \frac{1}{|\mathscr{A}|}$

Else # $\Pr(X = \text{greedy}) = 1 - \epsilon$

$A_t \sim U(\mathscr{A}^*)$ # $\Pr(A_t = a \,|\, X = \text{greedy}) = \mathbf{1}_{a \in \mathscr{A}^*} \frac{1}{|\mathscr{A}^*|}$

PROBABILITIES

ACTION SELECTION

$\epsilon -$Greedy action selection

$$\Pr(A_t = a) = \Pr\left(\left(A_t = a, X = \text{random}\right) \cup \left(A_t = a, X = \text{greedy}\right)\right)$$

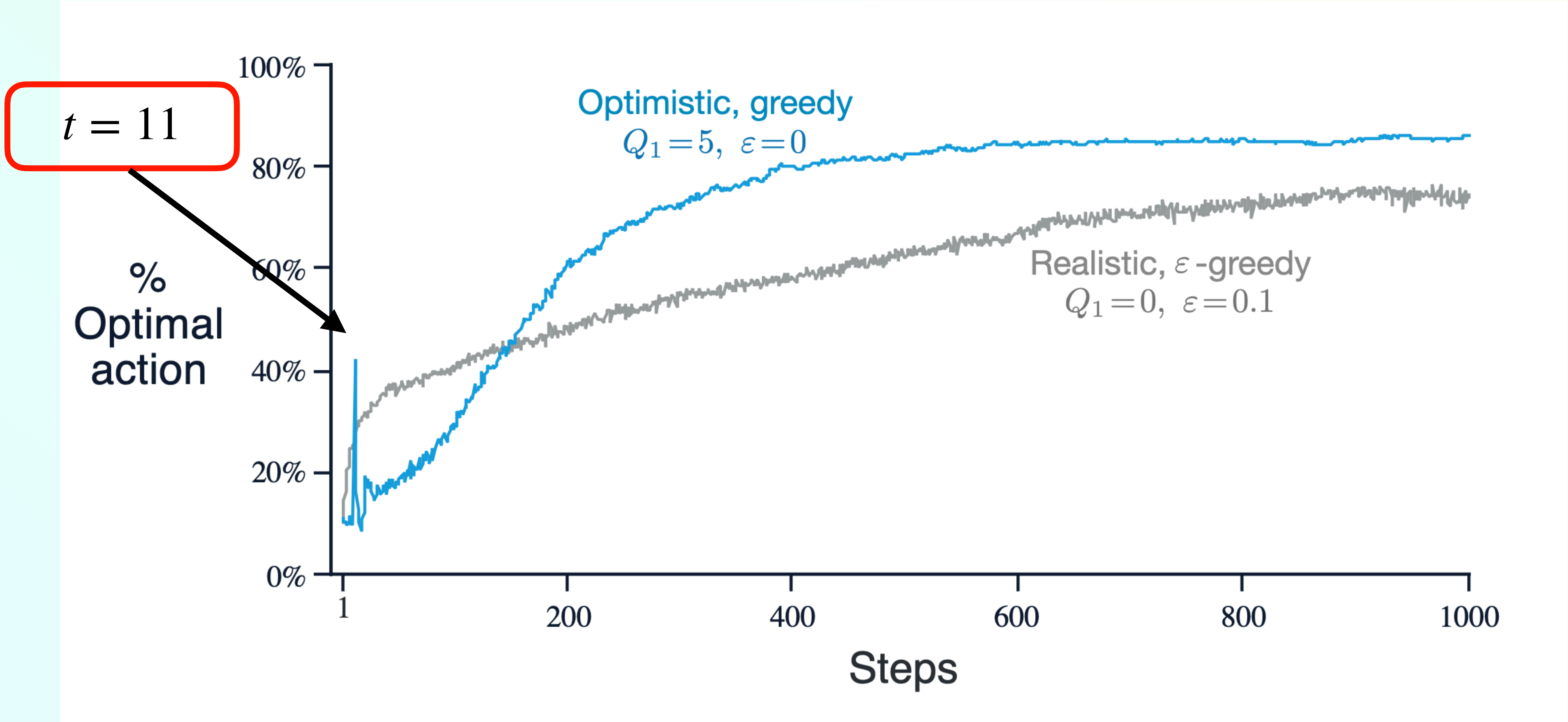
$$\Pr(A_t = a, X = \text{random}) = \Pr\left(A_t = a \,|\, X = \text{random}\right) \Pr\left(X = \text{random}\right) = \frac{1}{|\mathscr{A}|}\epsilon$$

$$\Pr(A_t = a, X = \text{greedy}) = \Pr\left(A_t = a \,|\, X = \text{greedy}\right) \Pr\left(X = \text{greedy}\right) = \mathbf{1}_{a \in \mathscr{A}*}\frac{1}{|\mathscr{A}*|}(1 - \epsilon)$$

$$\Pr(A_t = a) = \frac{\epsilon}{|\mathscr{A}|} + \mathbf{1}_{a \in \mathscr{A}*}\frac{1 - \epsilon}{|\mathscr{A}*|}$$

OPTIMISTIC INITIALIZATION

SPIKES IN LEARNING



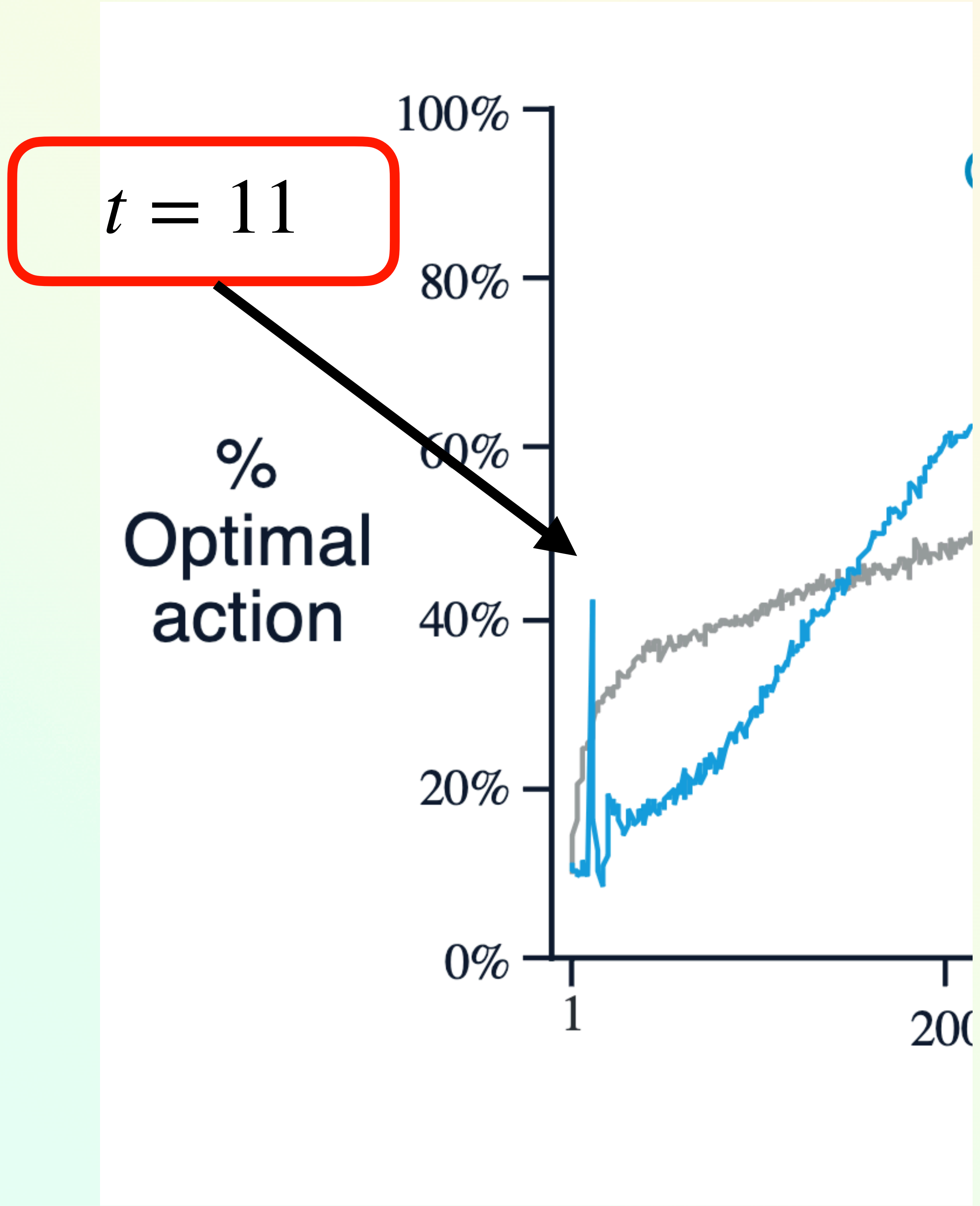
OPTIMISTIC INITIALIZATION

SPIKES IN LEARNING

What keeps performance low $t \in [1{,}10]$?

Why is there a spike at $t = 11$?

Why does performance drop at $t = 12$?



OPTIMISTIC INITIALIZATION

SPIKES IN LEARNING

What keeps performance low $t \in [1,10]$?

- $\forall a \; Q_1(a) = 5$
- $Q_{n+1} = Q_n + \alpha \left(R_n - Q_n\right)$
- $R_n - Q_n < 0 \; -> Q_{n+1}$ will be smaller
- Each action will be tried in the first 10 steps

t	Q(1)	Q(2)	Q(3)	Q(4)	A_t	R_t
1	**5**	**5**	**5**	**5**	1	2
2	4.7	**5**	**5**	**5**	3	3
3	4.7	**5**	4.8	**5**	4	1
4	4.7	**5**	4.8	4.6	2	0
10	4.6	4.5	4.8	4.6	Last unchosen action	0

OPTIMISTIC INITIALIZATION

SPIKES IN LEARNING

Why is there a spike at $t = 11$?

- The first action that is influenced by rewards

$$\arg\max_a Q_2(a) = \arg\max_a (1-\alpha)Q_1(a) + \alpha R_1^{(a)} = \arg\max_a (1-\alpha)5 + \alpha R_1^{(a)}$$

- $$= \arg\max_a R_1^{(a)}$$

- A_{11} more likely to be optimal than A_{10}

t	Q(1)	Q(2)	Q(3)	Q(4)	A_t	R_t
11	4.6	4.5	**4.8**	4.6	3	

OPTIMISTIC INITIALIZATION

SPIKES IN LEARNING

Why does performance drop at $t = 12$?

- $R_2^{A_{11}} - Q_2(A_{11}) < 0 \quad Q$ decreases
- Choose the second-best action from $t = 11$
- Value keeps decreasing

t	Q(1)	Q(2)	Q(3)	Q(4)	A_t	R_t
11	4.6	4.5	**4.8**	4.6	3	2
12	**4.6**	4.5	4.52	**4.6**	1	3
13	4.44	4.5	4.52	**4.6**	4	2
14	4.44	4.5	**4.52**	4.34	3	3
15	4.44	**4.5**	4.368	4.34	2	

CODE EXAMPLES

INSTALLING JULIA AND PLUTO

Download and install Julia: https://julialang.org/downloads/

Install Pluto

1. Run Julia
2.
```
julia> using Pkg

julia> Pkg.add("Pluto")
```

Guide: https://computationalthinking.mit.edu/Spring21/installation/

CODE EXAMPLES

LAUNCHING A PLUTO NOTEBOOK

1. Download notebook
2. Launch Pluto

```
julia> using Pluto

julia> Pluto.run()
```

3. Select notebook file

CODE EXAMPLES

LAUNCHING A PLUTO NOTEBOOK

See notebook

UPPER CONFIDENCE BOUND (UCB)

MODELING UNCERTAINTY

Greedy fails because we trust the estimates completely

- Assume $\forall a \ Q_n(a) = q_*(a)$ – Only true at $n \to \infty$

Model uncertainty on an estimate of $q_*(a)$ as confidence interval

UPPER CONFIDENCE BOUND (UCB)

CONFIDENCE INTERVALS

$\mu = \mathbb{E}[X]$ – parameter we want to know

$\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ – set of n observations

$L(\mathcal{X}_n)$ – a function that computes a lower bound on μ given the data

$U(\mathcal{X}_n)$ – a function that computes an upper bound on μ given the data

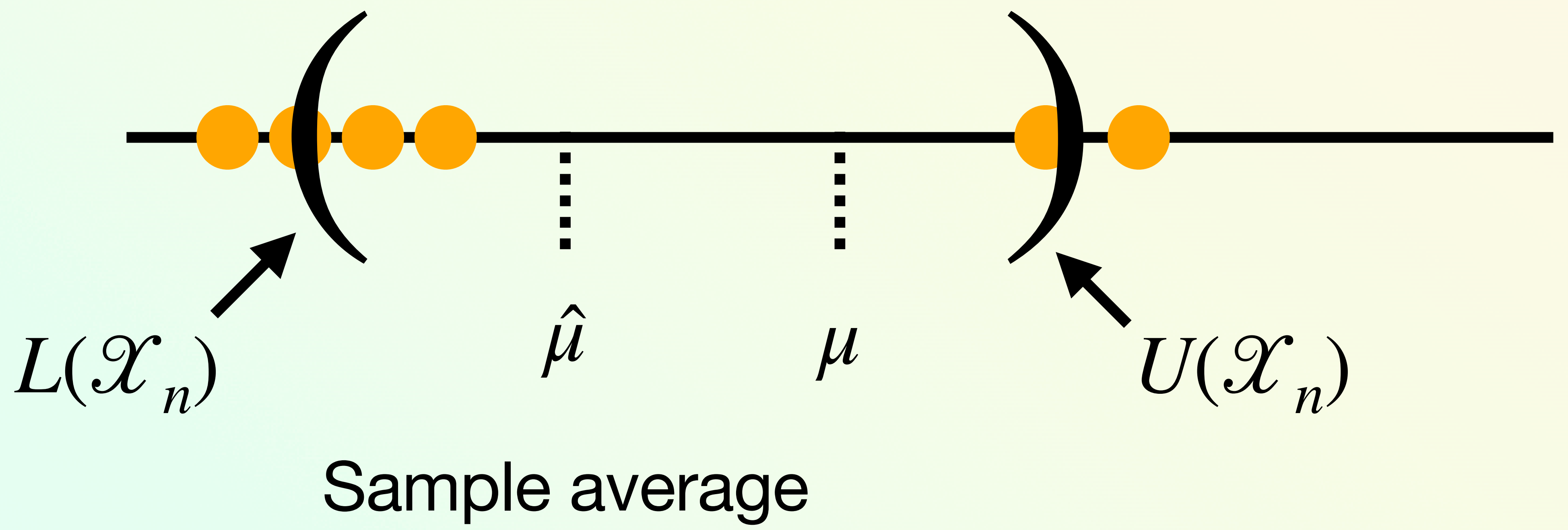
$\delta \in (0,1)$ – confidence level, saying how often the interval can fail

$$\Pr\left(\mu \in \left[L(\mathcal{X}_n), U(\mathcal{X}_n)\right]\right) \geq 1 - \delta$$

UPPER CONFIDENCE BOUND (UCB)

CONFIDENCE INTERVALS

$$\Pr\left(\mu \in \left[L(\mathcal{X}_n), U(\mathcal{X}_n)\right]\right) \geq 1 - \delta$$



$\delta = 0.05$ —> $95\,\%$ confidence intervals

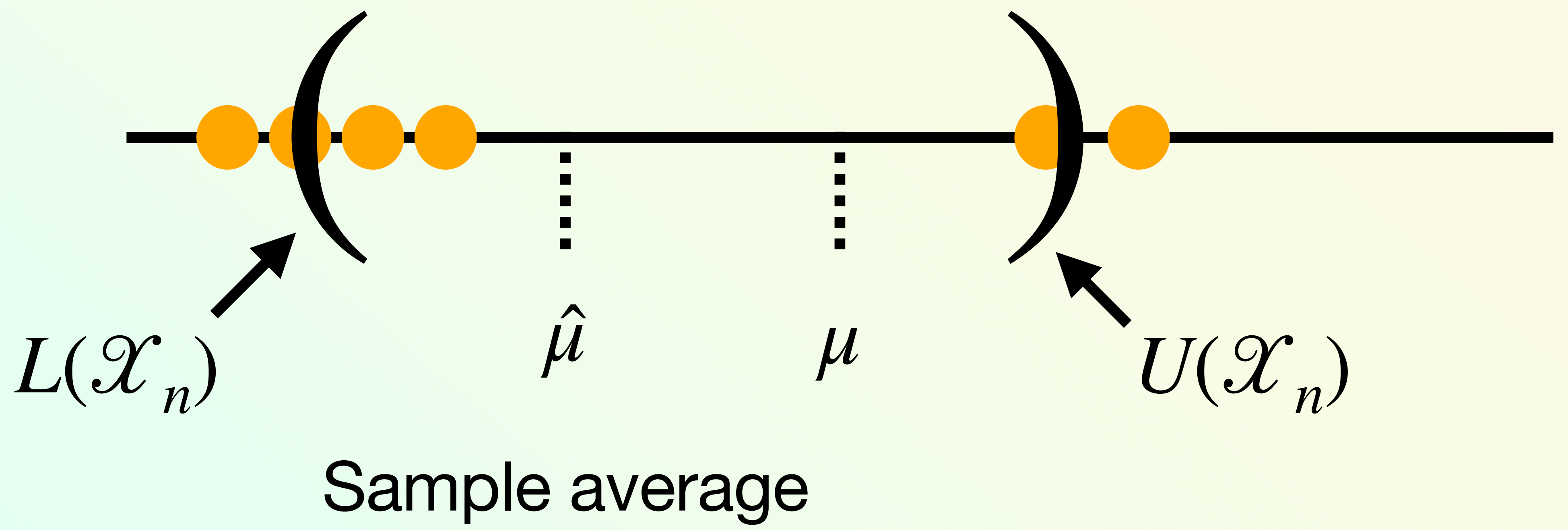
Confidence interval says that the construction of the interval will contain the mean with probability $1 - \delta$

Confidence intervals do not imply that the mean is in the interval with probability $1 - \delta$

UPPER CONFIDENCE BOUND (UCB)

CONFIDENCE INTERVALS

$$\Pr\left(\mu \in \left[L(\mathscr{X}_n), U(\mathscr{X}_n)\right]\right) \geq 1 - \delta$$



$\delta = 0.05$ –> $95\ \%$ confidence intervals

Confidence interval says that the construction of the interval will contain the mean with probability $1 - \delta$

Confidence intervals do not imply that the mean is in the interval with probability $1 - \delta$

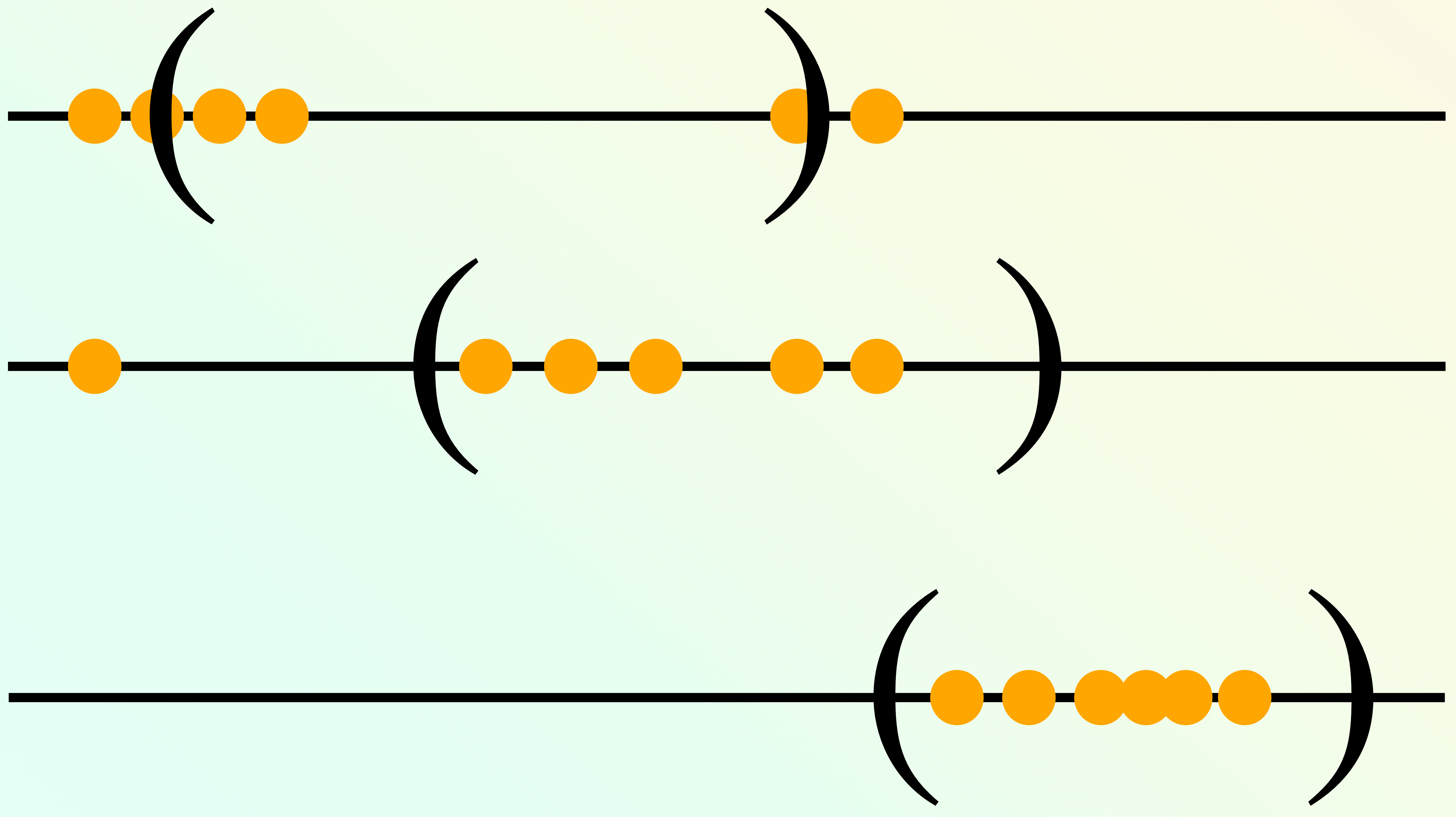
UPPER CONFIDENCE BOUND (UCB)

CONFIDENCE INTERVALS: COMPARISION

If $U(\mathscr{X}) < L(\mathscr{Y})$

$\mu_X < \mu_Y$ with confidence $1 - 2\delta$

If intervals overlap, we cannot tell if the means are different



UPPER CONFIDENCE BOUND (UCB)

CONFIDENCE INTERVALS METHODS

Students t-distribution interval

$\hat{\mu}(\mathcal{X}_n) = \frac{1}{n}\sum_{i=1}^{n} X_1$ Sample mean

$\hat{\sigma}(\mathcal{X}_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_1 - \hat{\mu}(\mathcal{X}_n)\right)^2}$ Sample standard deviation

$t_{1-\delta,v}$ $100(1-\delta)$ percentile of the Student t-distribution with v degrees of freedom

$\hat{\mu}(\mathcal{X}_n) \pm t_{1-\delta,n-1}\frac{\hat{\sigma}(\mathcal{X}_n)}{\sqrt{n}}$ The confidence interval centered around the sample mean

It is more likely to produce a valid confidence interval as $n \to \infty$

- Usually needs at least 30 samples

UPPER CONFIDENCE BOUND (UCB)

CONFIDENCE INTERVALS METHODS

Confidence interval based on Hoeffding's inequality

Requires: $\forall i,\ X_i \in [a, b]$

$$\hat{\mu}(\mathcal{X}_n) \pm (b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}$$

Valid confidence interval for all $n \geq 1$

The interval is very wide and needs lots of data to detect differences

UPPER CONFIDENCE BOUND (UCB)

UCB

Select the action that *might* have the highest value

Uncertainty decreases as we sample an action so we can rule out some bad actions

Select actions greedily from upper bound

$$A_t \in \text{arg}\max_a Q_t(a) + c\sqrt{ln(t)/N_t(a)}$$

$N_t(a)$ number of times a was chosen up until time t

If $N_t(a) = 0$ then the action is treated as having the highest upper bound

UPPER CONFIDENCE BOUND (UCB)

UCB

$$A_t \in \text{arg}\max_a Q_t(a) + c\sqrt{ln(t)/N_t(a)}$$

The upper bound increases if the action is not chosen

c needs to be large enough to make sure the upper bound is not too low

UPPER CONFIDENCE BOUND (UCB)

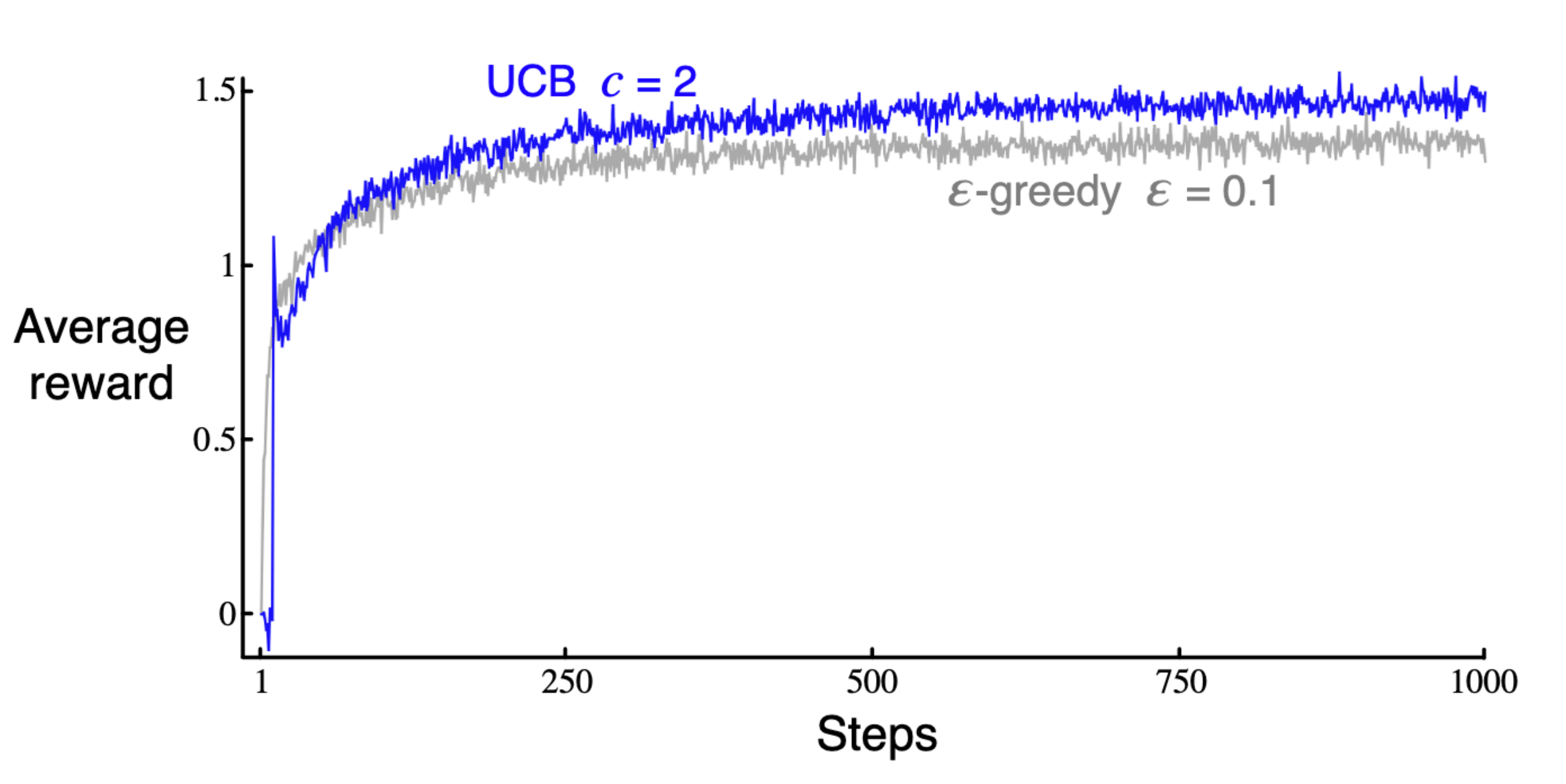
UCB CODE EXAMPLE

See notebook

UPPER CONFIDENCE BOUND (UCB)

UCB QUESTION

Why is there a spike at $t = 11$?



SOFTMAX

ACTION SELECTION BASED ON Q VALUES

$\epsilon-$ Greedy

- Samples the same action often even if there are other good ones to learn about
- treats all non-greedy actions the same
- Small change in Q can lead to a big change in which actions are being chosen

Idea: Sample actions relative to the value of the action

SOFTMAX

ACTION SELECTION BASED ON Q VALUES

$$\Pr(A_t = a) = \frac{e^{\tau Q_t(a)}}{\sum_{b \in \mathcal{A}} e^{\tau Q_t(b)}}$$

- Small estimates will have a low chance of being chosen
- Large estimates will have a high chance of being chosen
- τ — temperature parameter
  - $\tau \to 0$ distribution becomes uniform
  - $\tau \to \infty$ distribution becomes greedy

HOW MUCH EXPLORATION

DIFFERENT REQUIREMENTS

Infinite lifetime:

- Exploration needs to decrease with time

Limited Lifetime:

- Strictly balance between exploration and exploitation based on time remaining

Nonstationary:

- The agent needs to retry actions that were bad before (they might be good now)

NEXT CLASS

WHAT YOU SHOULD DO

1. Programming assignment due tonight
2. Watch week 2 videos on MDPs before next class
3. Quiz due Friday night

Friday: MDP overview