

“And always, he fought the temptation to choose a clear, safe course, warning “That path leads ever down into stagnation.””

Frank Herbert, *Dune*



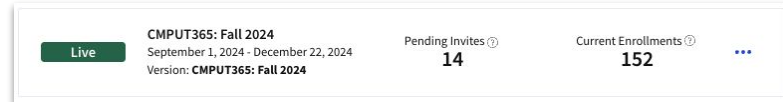
CMPUT 365
Introduction to
Sequential-Decision Making

Plan

- Wrap up reviewing content
- Exercises!

Reminder

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.



Live	CMPUT365: Fall 2024 September 1, 2024 - December 22, 2024 Version: CMPUT365: Fall 2024	Pending Invites ⓘ 14	Current Enrollments ⓘ 152	⋮
------	--	--------------------------------	-------------------------------------	---

I **cannot** use marks from the public repository for your course marks.

You **need to check, every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

The deadlines in the public session **do not align** with the deadlines in Coursera.

If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

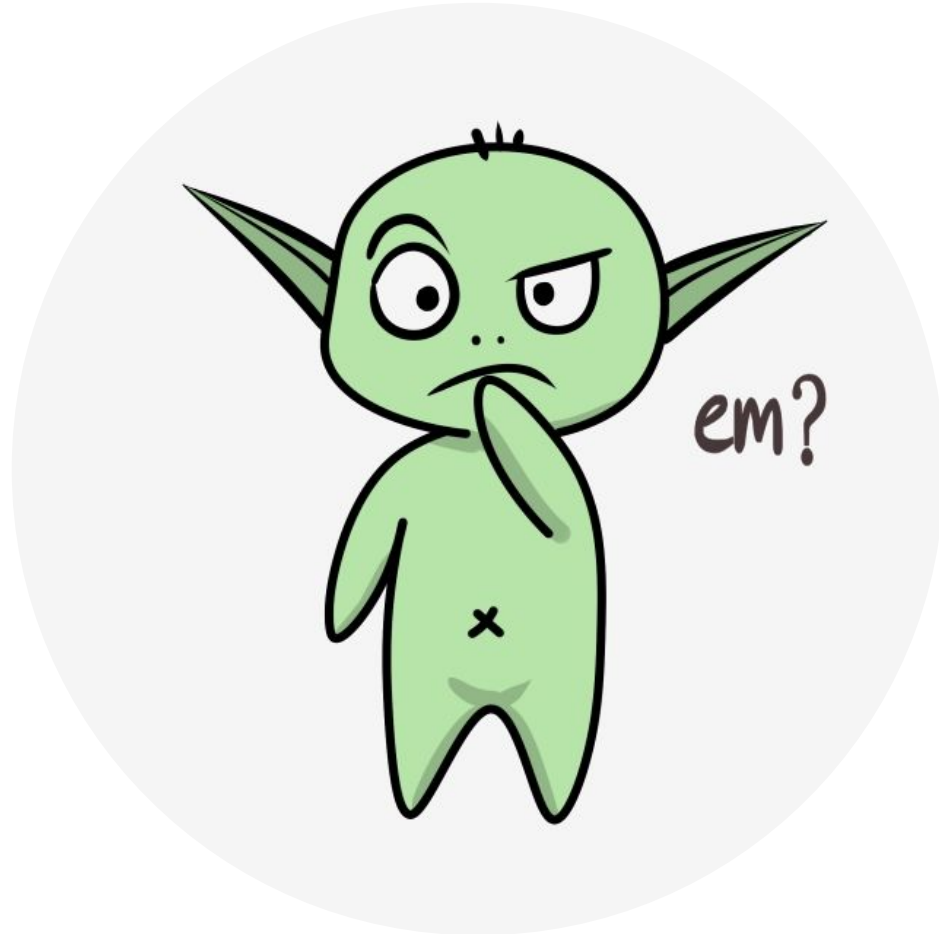
Please, interrupt me at any time!



Last Class: Incremental updates to estimate q_*

$$\begin{aligned} Q_{n+1} &= \frac{1}{n} \sum_{i=1}^n R_i \\ &= Q_n + \frac{1}{n} [R_n - Q_n], \end{aligned}$$

NewEstimate \leftarrow OldEstimate + StepSize [Target - OldEstimate]



Update rule

$$\text{NewEstimate} \leftarrow \text{OldEstimate} + \text{StepSize} [\text{Target} - \text{OldEstimate}]$$

$$Q_{n+1} \doteq Q_n + \alpha [R_n - Q_n]$$

A bigger step-size means bigger steps (updates).

A constant step-size gives more weight to recent rewards.

How you initialize Q_n really matters.

The principle of **optimism in the face of uncertainty** really leverages that.

This is the direction you need to move to get closer to the solution.

A note on step-sizes

A well-known result in stochastic approximation theory gives us the conditions required to assure convergence with probability 1:

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty$$

Cannot be too small.
E.g.: $\alpha_n = 1/n^2$

and

$$\sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

Cannot be too big.
E.g.: $\alpha_n = 1$

A constant step-size is biased

$$Q_{n+1} = Q_n + \alpha [R_n - Q_n]$$

A constant step-size is biased

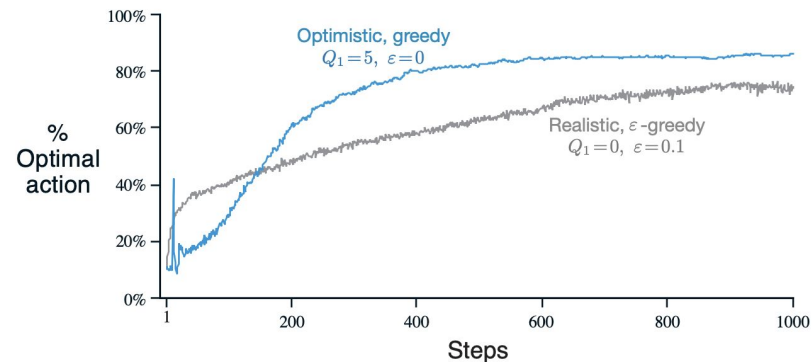
$$\begin{aligned}Q_{n+1} &= Q_n + \alpha [R_n - Q_n] \\&= \alpha R_n + (1 - \alpha) Q_n \\&= \alpha R_n + (1 - \alpha) [\alpha R_{n-1} + (1 - \alpha) Q_{n-1}] \\&= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\&= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \\&\quad \dots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\&= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i.\end{aligned}$$

Q_1 is always there, forever,
impacting the final estimate.

Optimism in the face of uncertainty

$$Q_{n+1} = Q_n + \alpha [R_n - Q_n]$$

Idea: Initialize Q_0 to an overestimation of its true value (optimistically).

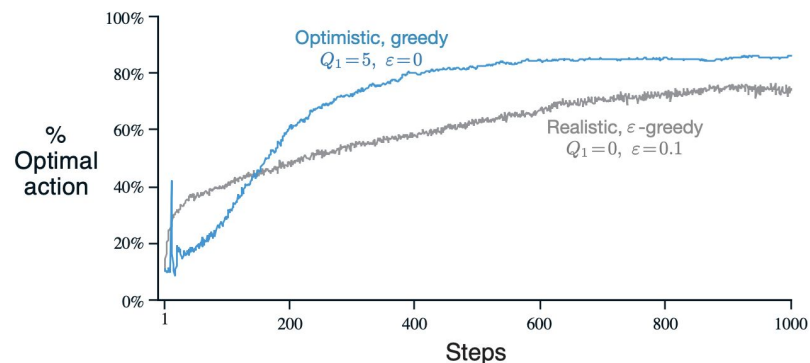


Optimism in the face of uncertainty

$$Q_{n+1} = Q_n + \alpha [R_n - Q_n]$$

Idea: Initialize Q_0 to an overestimation of its true value (optimistically).

- You either maximize reward or you learn from it.
- The value you initialize Q_0 can be seen as a hyperparameter and it matters.
- There are equivalent transformations in the reward signal to get the same effect.
- For bandits, UCB uses an upper confidence bound that with high probability is an overestimate of the unknown value.



How do we choose the best hyperparameter (α , ϵ , c , etc)?

- For this course: we try many things out and see what works best $_(\ツ)_/$





Upper-Confidence-Bound Action Selection

$$A_t \doteq \arg \max_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

Theorem 1. *For all $K > 1$, if policy UCB1 is run on K machines having arbitrary reward distributions P_1, \dots, P_K with support in $[0, 1]$, then its expected regret after any number n of plays is at most*

$$\left[8 \sum_{i: \mu_i < \mu^*} \left(\frac{\ln n}{\Delta_i} \right) \right] + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{j=1}^K \Delta_j \right)$$

where μ_1, \dots, μ_K are the expected values of P_1, \dots, P_K .

Auer, Cesa-Bianchi, and Fischer (2002), *Machine Learning*.

Contextual bandits (Associative search)

- One need to associate different actions with different *situations*.
- You need to learn a *policy*, which is a function that maps situations to actions.
- Most real-world problems modeled as bandits problems are modeled as contextual bandits problems.
- Example: A recommendation system, which is obviously conditioned on the user to which the system is making recommendations to.



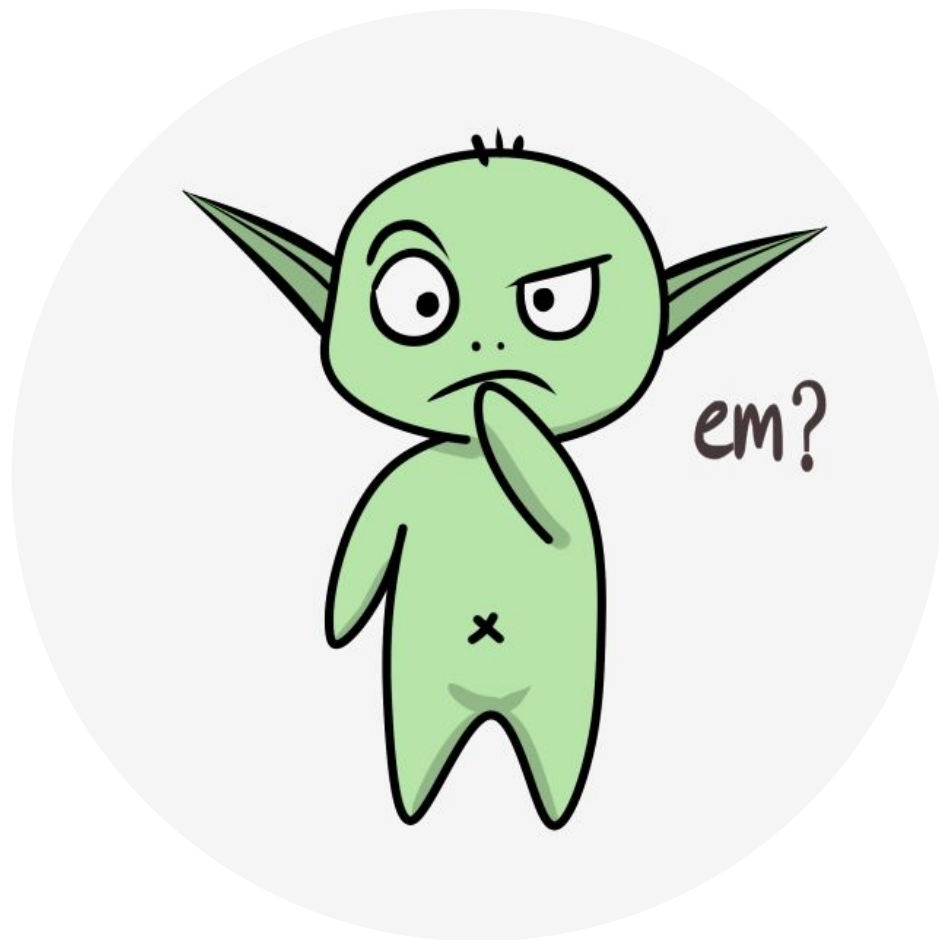
Exercise – Modeling

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

Suppose a game where you choose to flip one of two (possibly unfair) coins. You win \$1 if your chosen coin shows heads and lose \$1 if it shows tails.

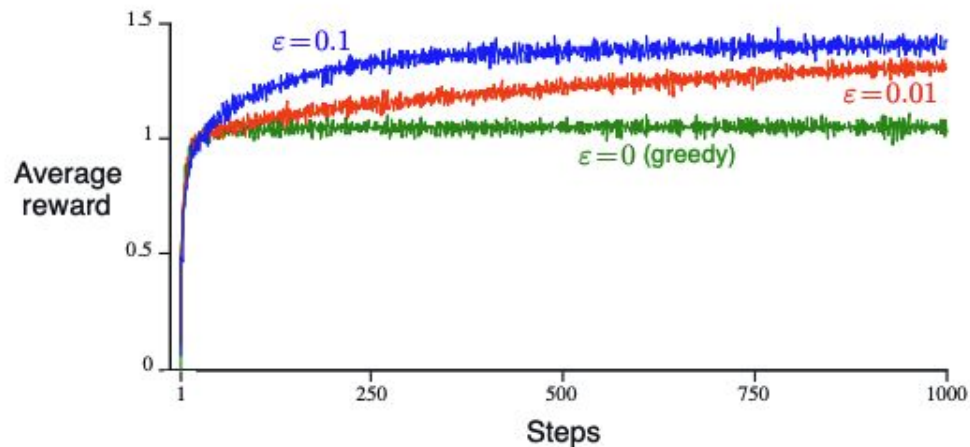
1. Model this as a K-armed bandit problem: define the action set.
2. Is the reward a deterministic or stochastic function of your action?
3. You do not know the coin flip probabilities. Instead, you are able to view 6 sample flips for each coin respectively: (T, H, H, T, T, T) and (H, T, H, H, H, T). Use the sample average formula (equation 2.1 in the book) to compute the estimates of the value of each action.
4. Decide on which coin to flip next! Assume it's an exploit step.

Solution – Modeling

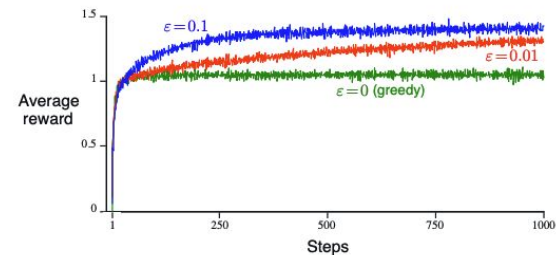


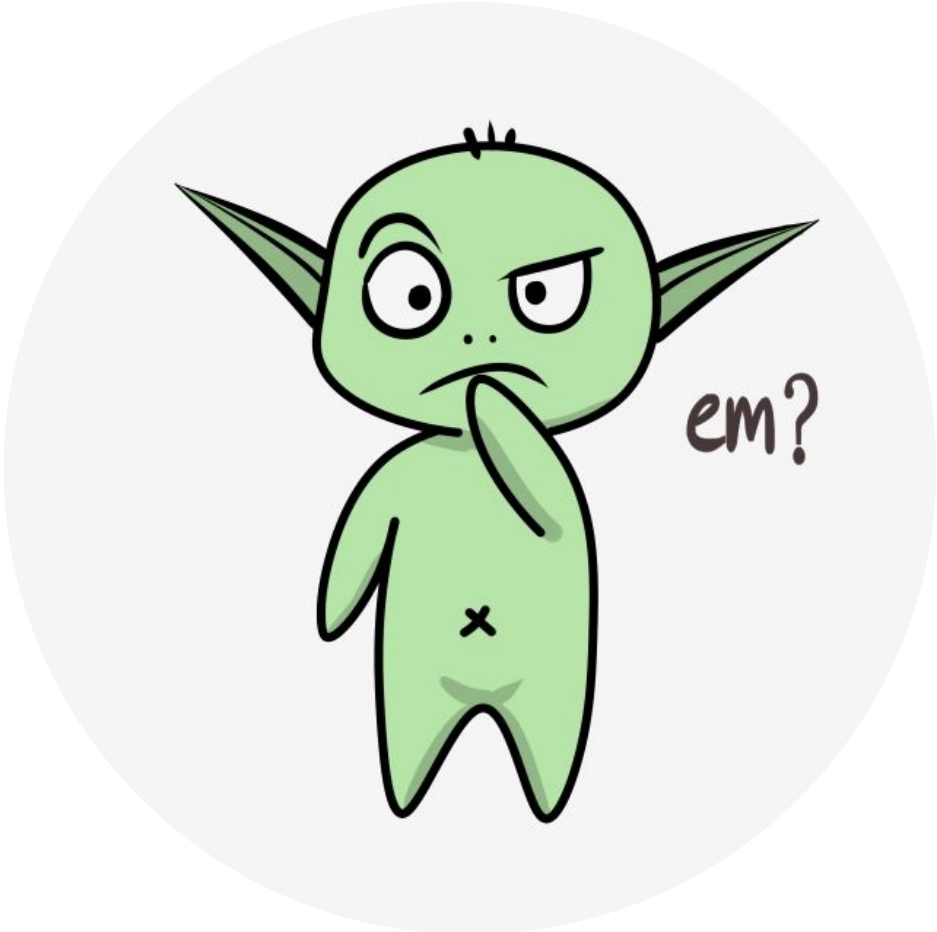
Exercise 2.3 of the textbook

Exercise 2.3 In the comparison shown in Figure 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answer quantitatively. \square



Solution – Exercise 2.3 of the textbook





Exercise 2.4 of the textbook

Exercise 2.4 If the step-size parameters, α_n , are not constant, then the estimate Q_n is a weighted average of previously received rewards with a weighting different from that given by (2.6). What is the weighting on each prior reward for the general case, analogous to (2.6), in terms of the sequence of step-size parameters? \square

Reminder:

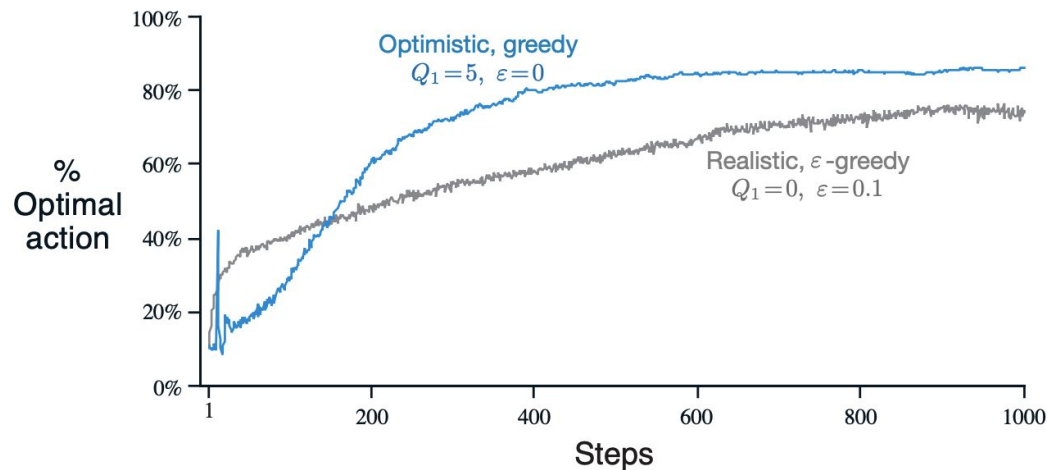
$$\begin{aligned}
 Q_{n+1} &= Q_n + \alpha [R_n - Q_n] \\
 &= \alpha R_n + (1 - \alpha) Q_n \\
 &= \alpha R_n + (1 - \alpha) [\alpha R_{n-1} + (1 - \alpha) Q_{n-1}] \\
 &= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 Q_{n-1} \\
 &= \alpha R_n + (1 - \alpha) \alpha R_{n-1} + (1 - \alpha)^2 \alpha R_{n-2} + \\
 &\quad \dots + (1 - \alpha)^{n-1} \alpha R_1 + (1 - \alpha)^n Q_1 \\
 &= (1 - \alpha)^n Q_1 + \sum_{i=1}^n \alpha (1 - \alpha)^{n-i} R_i.
 \end{aligned}$$

Solution – Exercise 2.4 of the textbook

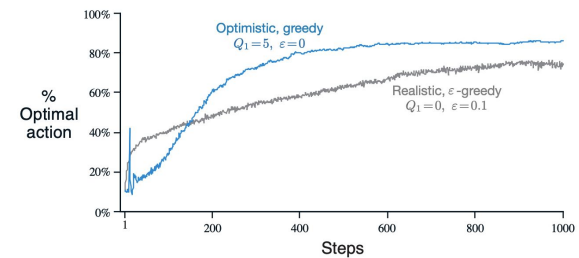


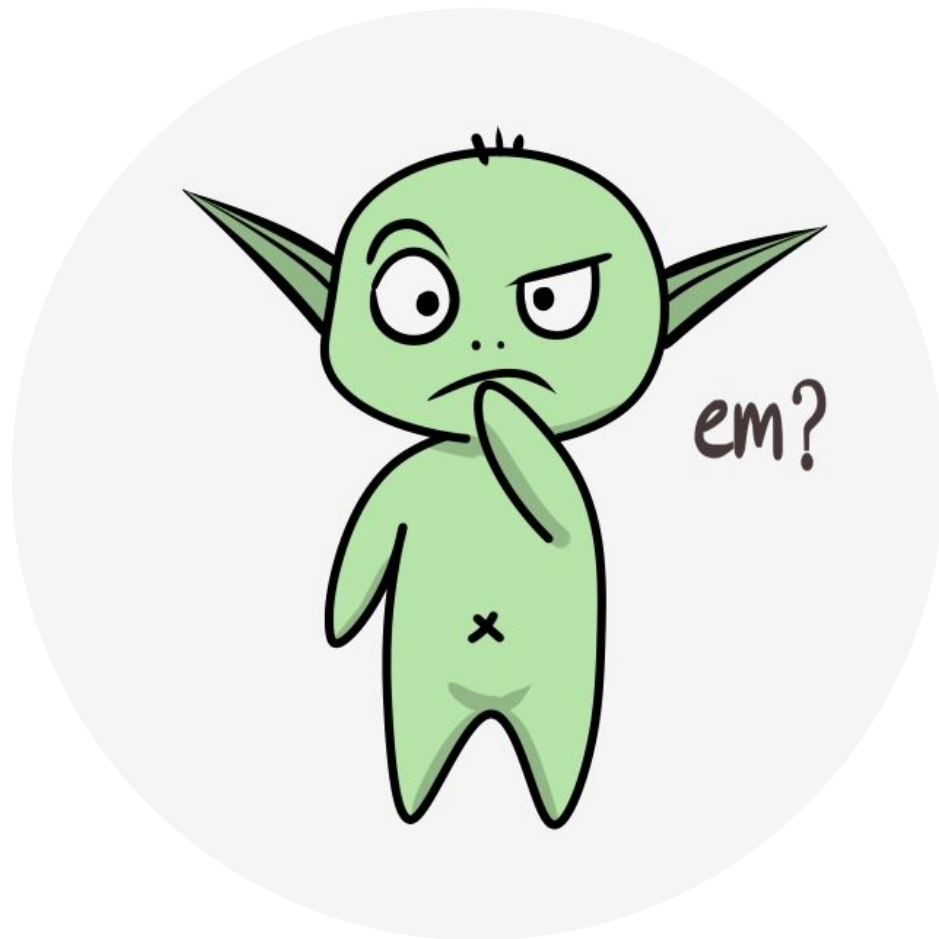
Exercise 2.6 of the textbook

Exercise 2.6: Mysterious Spikes The results shown in Figure 2.3 should be quite reliable because they are averages over 2000 individual, randomly chosen 10-armed bandit tasks. Why, then, are there oscillations and spikes in the early part of the curve for the optimistic method? In other words, what might make this method perform particularly better or worse, on average, on particular early steps? \square



Solution – Exercise 2.6 of the textbook





Exercise 2.7 of the textbook

Exercise 2.7: Unbiased Constant-Step-Size Trick In most of this chapter we have used sample averages to estimate action values because sample averages do not produce the initial bias that constant step sizes do (see the analysis leading to (2.6)). However, sample averages are not a completely satisfactory solution because they may perform poorly on nonstationary problems. Is it possible to avoid the bias of constant step sizes while retaining their advantages on nonstationary problems? One way is to use a step size of

$$\beta_n \doteq \alpha / \bar{o}_n, \tag{2.8}$$

to process the n th reward for a particular action, where $\alpha > 0$ is a conventional constant step size, and \bar{o}_n is a trace of one that starts at 0:

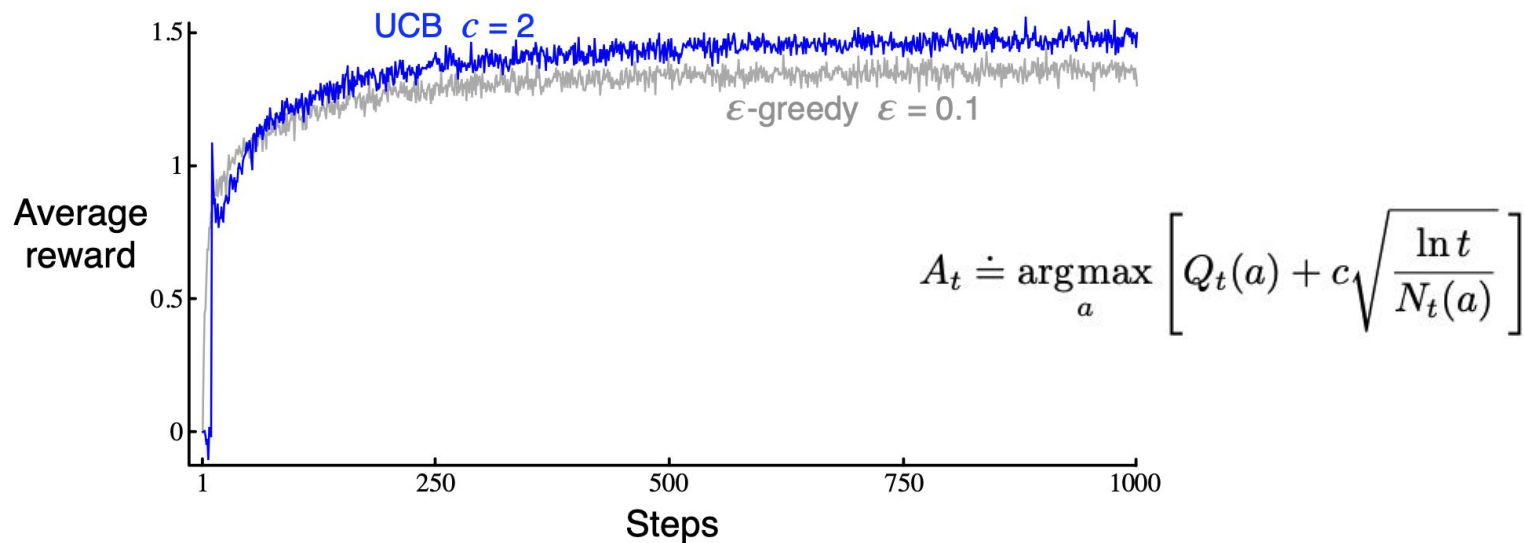
$$\bar{o}_n \doteq \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}), \text{ for } n > 0, \text{ with } \bar{o}_0 \doteq 0. \tag{2.9}$$

Carry out an analysis like that in (2.6) to show that Q_n is an exponential recency-weighted average *without initial bias*. \square

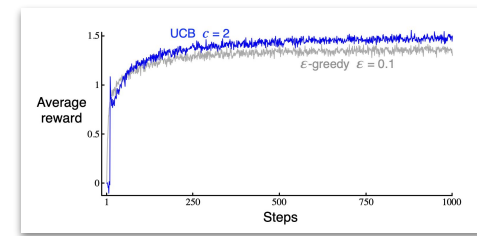
Solution – Exercise 2.7 of the textbook

Exercise 2.8 of the textbook

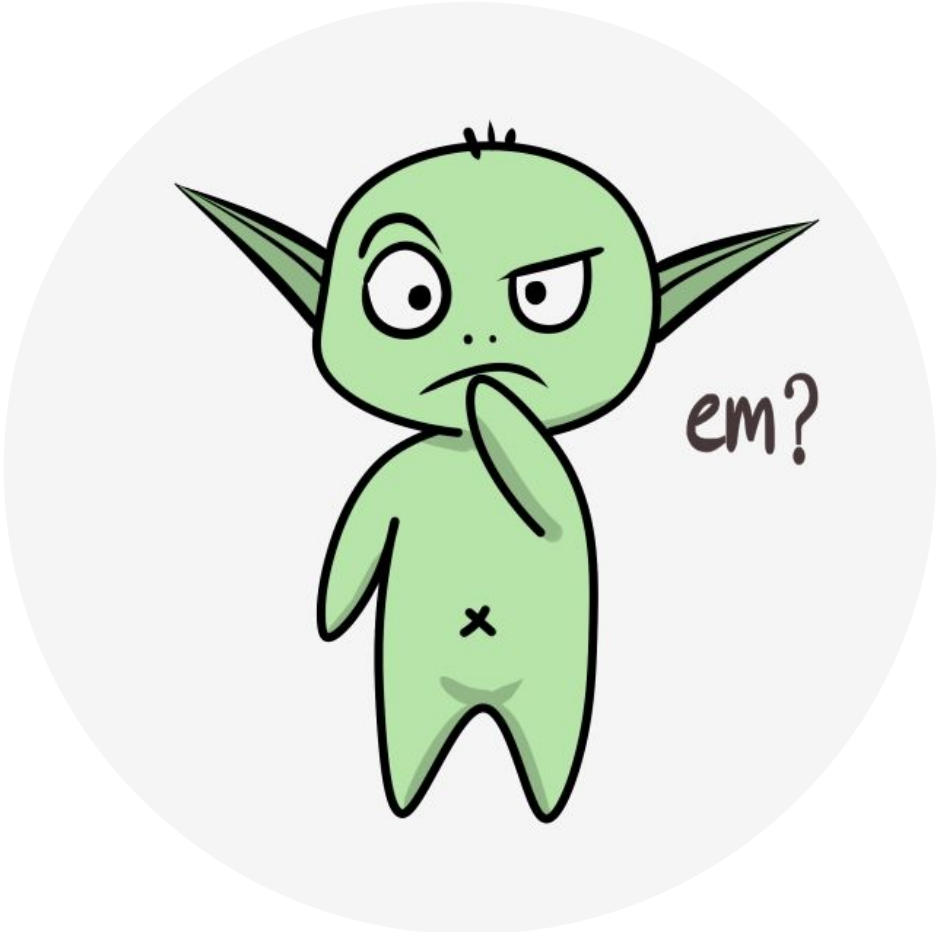
Exercise 2.8: UCB Spikes In Figure 2.4 the UCB algorithm shows a distinct spike in performance on the 11th step. Why is this? Note that for your answer to be fully satisfactory it must explain both why the reward increases on the 11th step and why it decreases on the subsequent steps. Hint: If $c = 1$, then the spike is less prominent. \square



Solution – Exercise 2.8 of the textbook



$$A_t \doteq \operatorname{argmax}_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$



Next class

Reminder: Programming Assignment for Coursera’s Fundamentals of RL: Sequential decision-making is due today at midnight.

- What **I** plan to do: Fundamentals of RL: Markov Decision Processes (MDPs)
 - Non-comprehensive overview about things related to MDPs (First half of chapter 3 of the textbook).
- What I recommend **YOU** to do for next class:
 - Watch videos of Week 2 of Coursera’s Fundamentals of RL (Module 1): M1W2.
 - Finish the recommended reading for Coursera’s M1W2.
 - Start collecting (and post) questions in eClass/Slack about the topic.
 - Submit practice quiz for Coursera’s Fundamentals of RL: MDPs (M1 W2).

You won’t have a graded assignment on Monday 😊