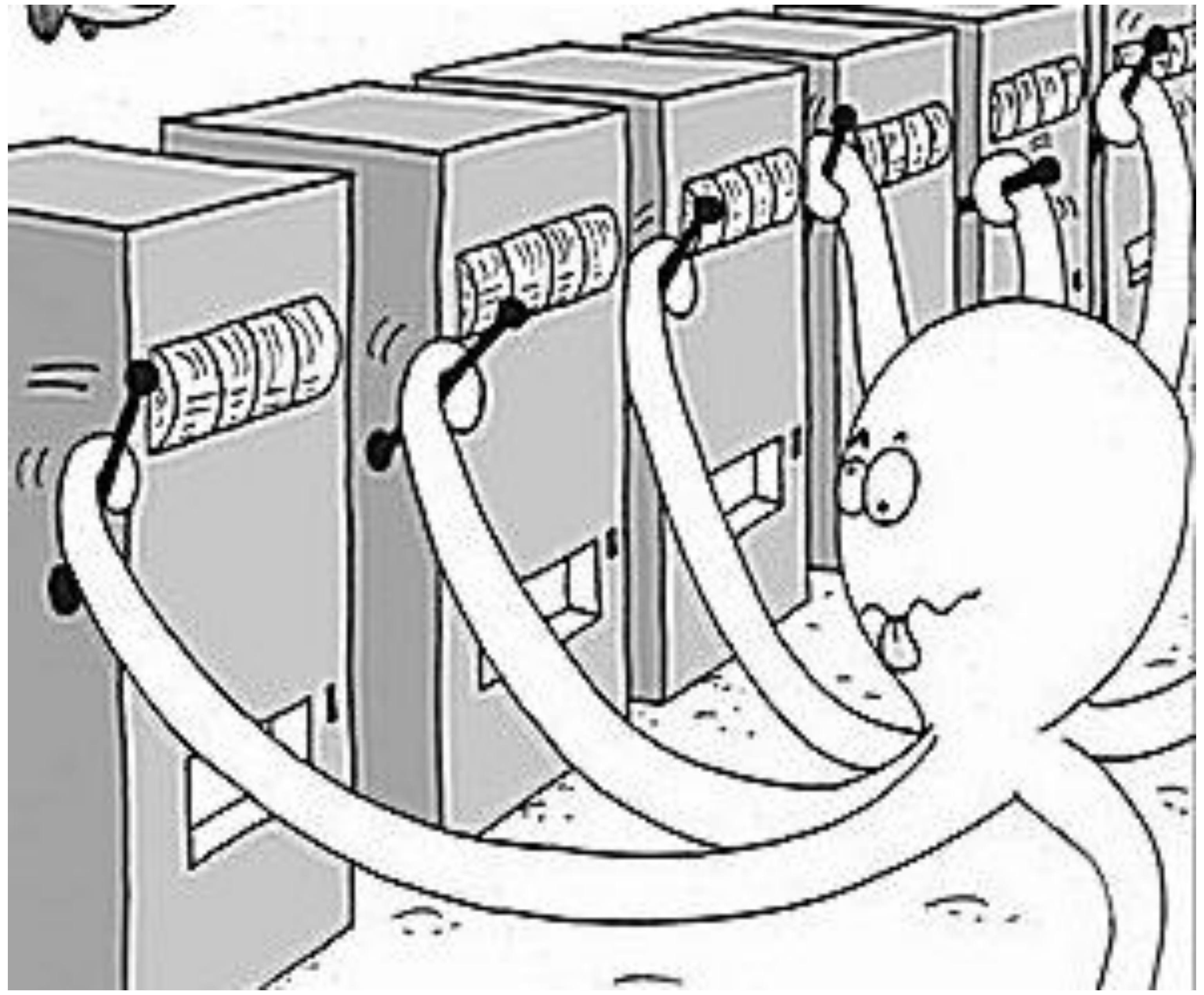
Work Session: seeing through the eyes of the agent

Course 1, Module 1
Sequential Decision Making

Reminders: Sept 8, 2021

- Schedule with deadlines on github pages (https://docs.google.com/spreadsheets/d/1ooFqttGCklw7rsst9xwL77_SA84LszLvZwpWo06Ltas)
- Next practice Quiz **due Sunday**, for Course 1, Module 2 (MDPs)
- *You all are reading along right??*
- TAs have posted office hours (likely different times next week). Over zoom/meet for now
- Any questions about admin?



Microsoft Research: http://slivkins.com/work/bandits-svc/

Demo of Bandits

- https://www.coursera.org/learn/fundamentals-of-reinforcement-learning/ungradedWidget/44Z9R/lets-play-a-game
- https://www.coursera.org/learn/fundamentals-of-reinforcement-learning/ungradedWidget/jEYTO/whats-underneath

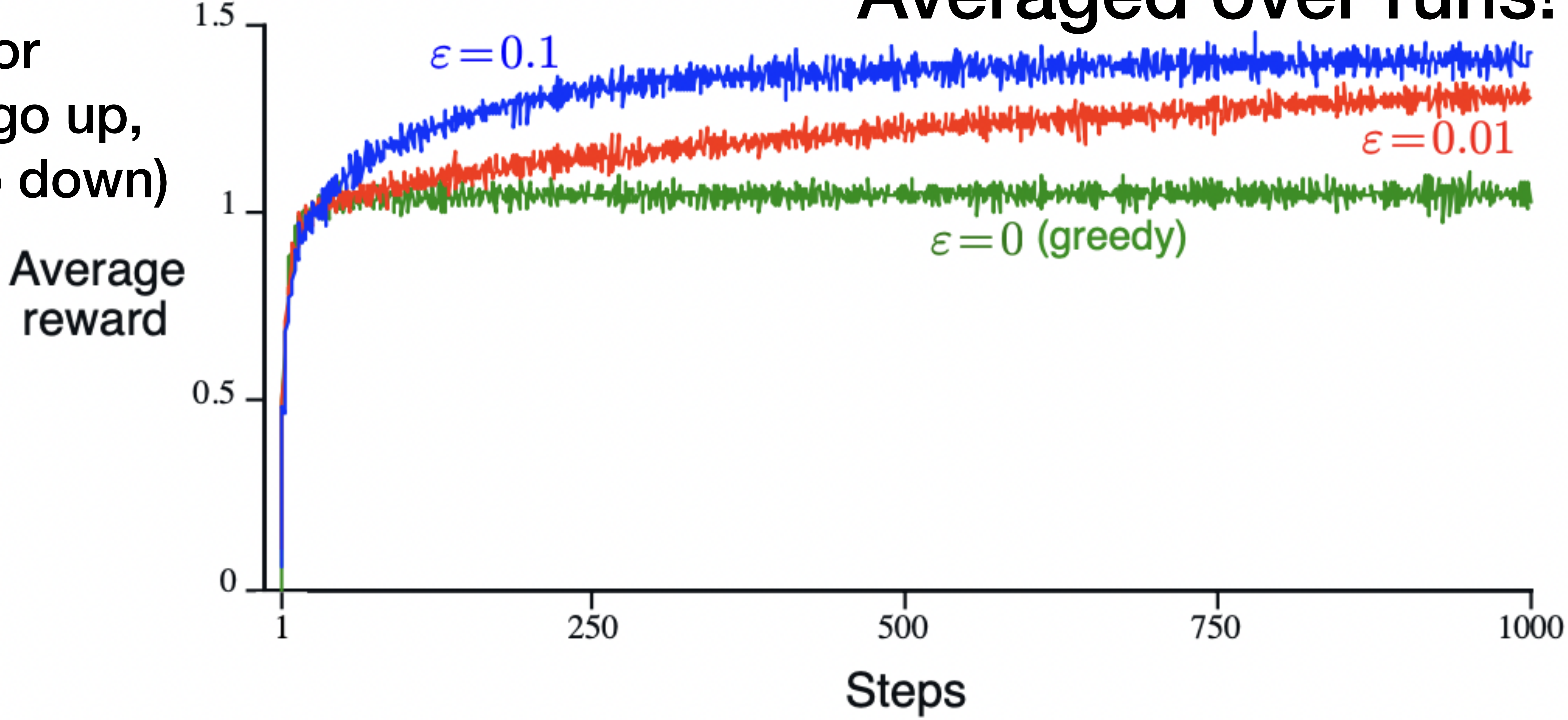
Digging deeper into learning curves

- There were some questions about impact of epsilon
- I noticed the curves in the notebook don’t match the book! Sup with that!?!
- Learning curves are super critical and we will be looking a lot at them, so comfort with them is key

Anatomy of a learning curve

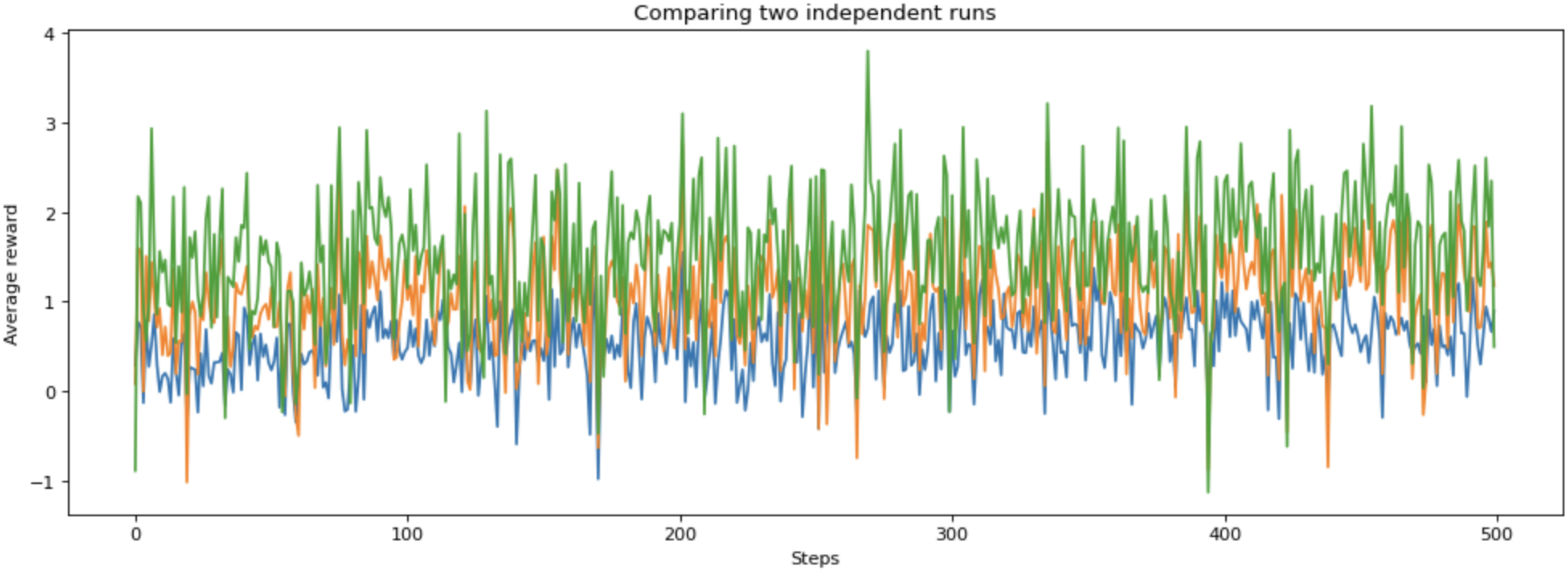
Performance
or
Error
(Some go up,
some go down)

Averaged over runs!



<-Experiment duration->

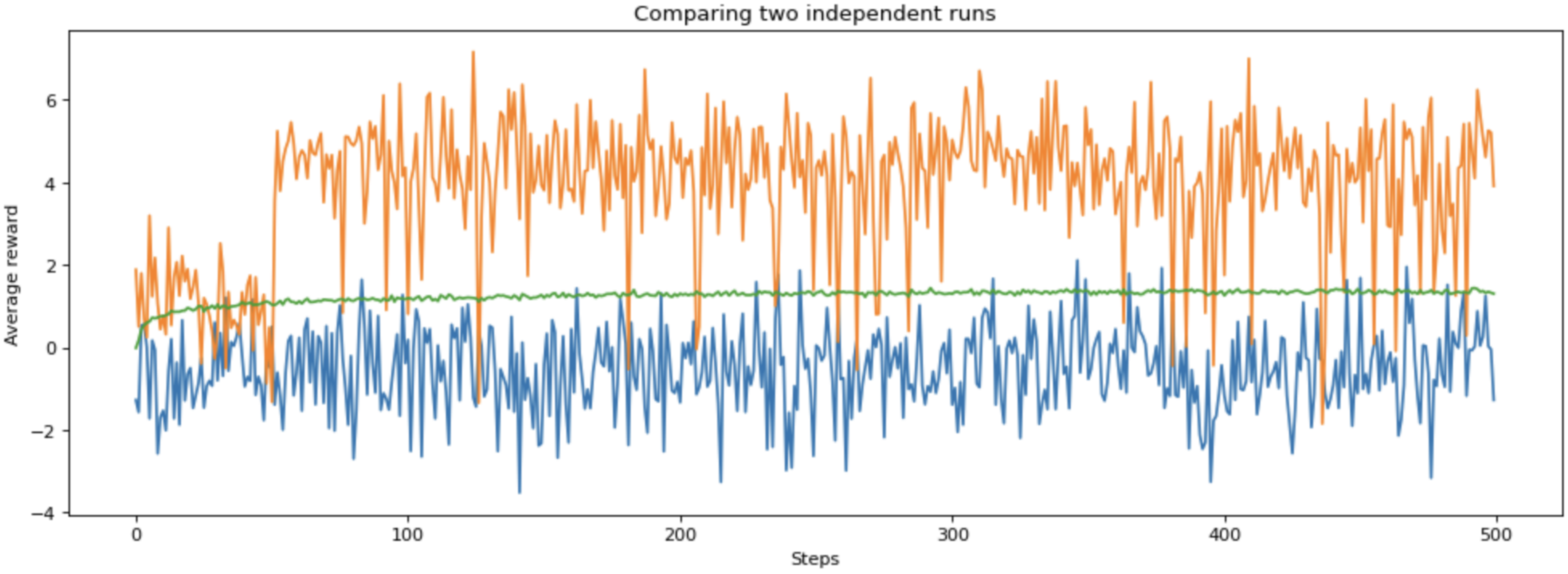
Individual runs of the system can be different due to randomness in rewards and the exploration



Averaging helps

- Imagine you wanted to empirically (using experiments and data analysis) convince me that the probability of rolling 2 on a fair six sided dice was 1/6?
- If you rolled it once that won't help
- If you rolled it 6 times you might never see a '2' **(unlucky)**
- If you rolled it 6 times, maybe you observe '2' five times **(lucky)**
- Using a large number of rolls and counting the '2's you observe, eventually will give you a good answer

Why we average

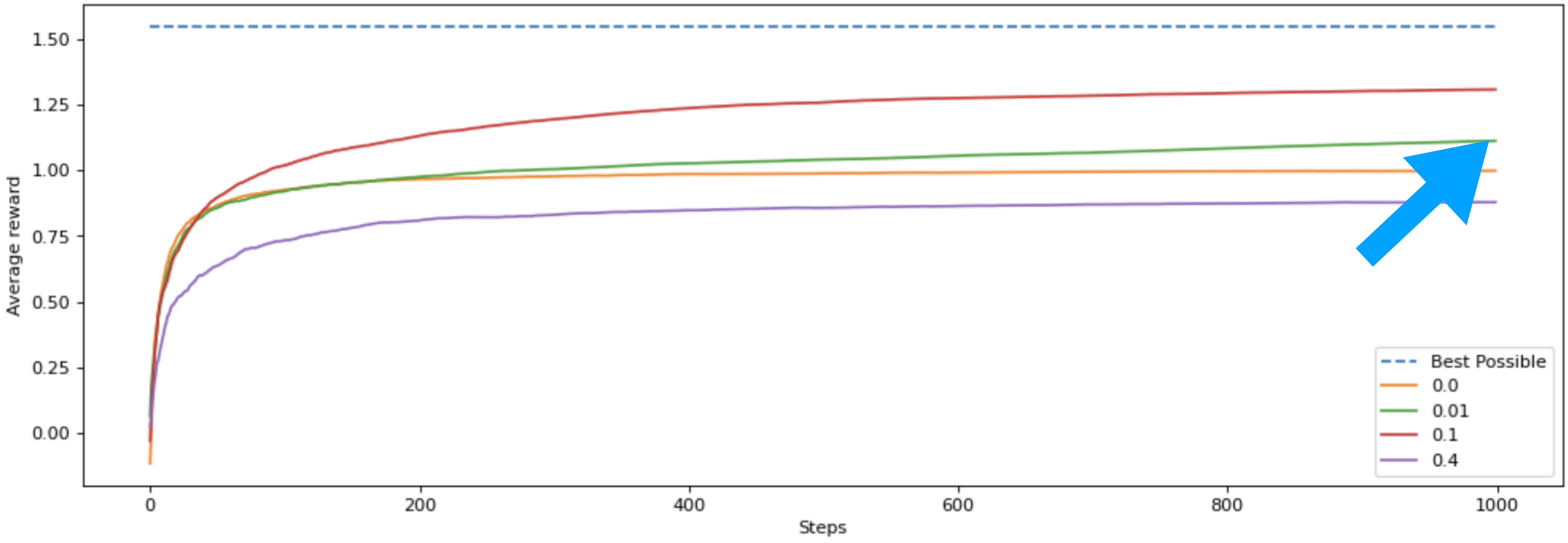


Agent might be lucky, or unlucky!

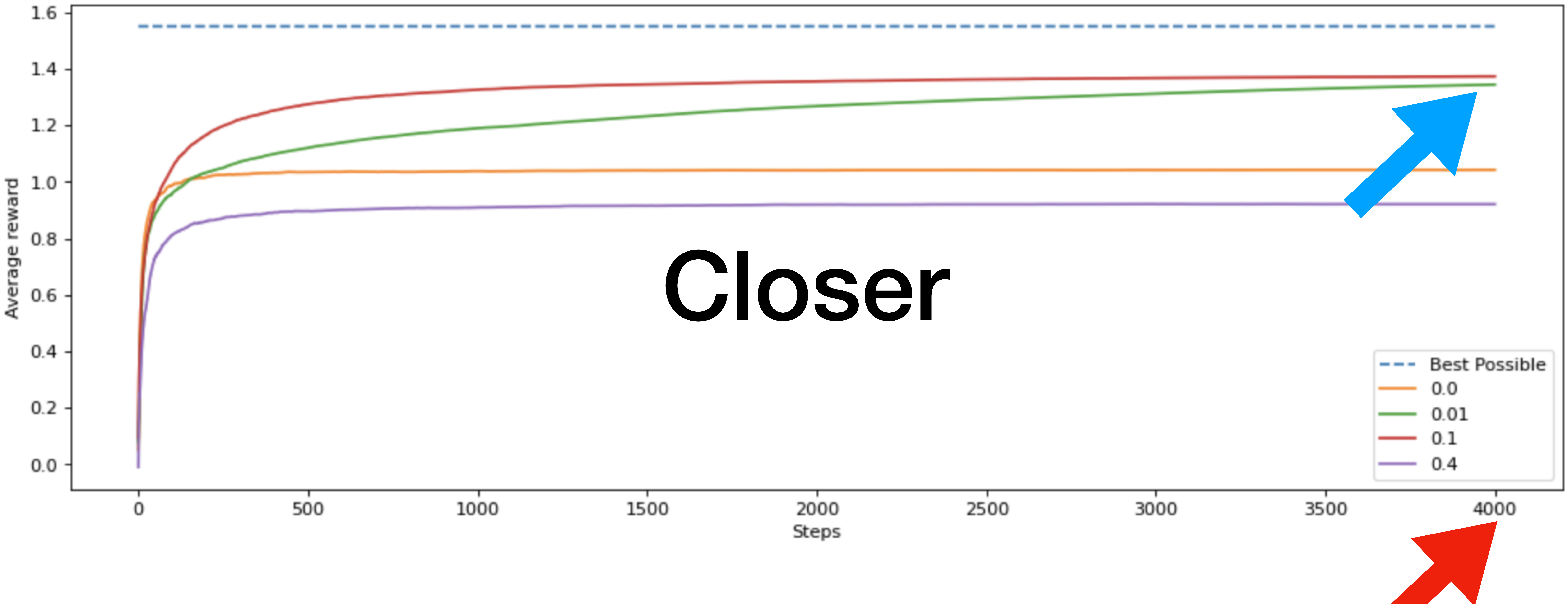
What else might we be interested in?

- Average or mean performance is useful in comparing algorithms
  - We can make confidence intervals around the mean, conduct hypothesis tests
  - Makes for clear and precise comparisons
- But in deployment we rarely care about average performance!
  - Average success in folding your laundry
- We could also report Best Performance
- We could also report Worst Case Performance

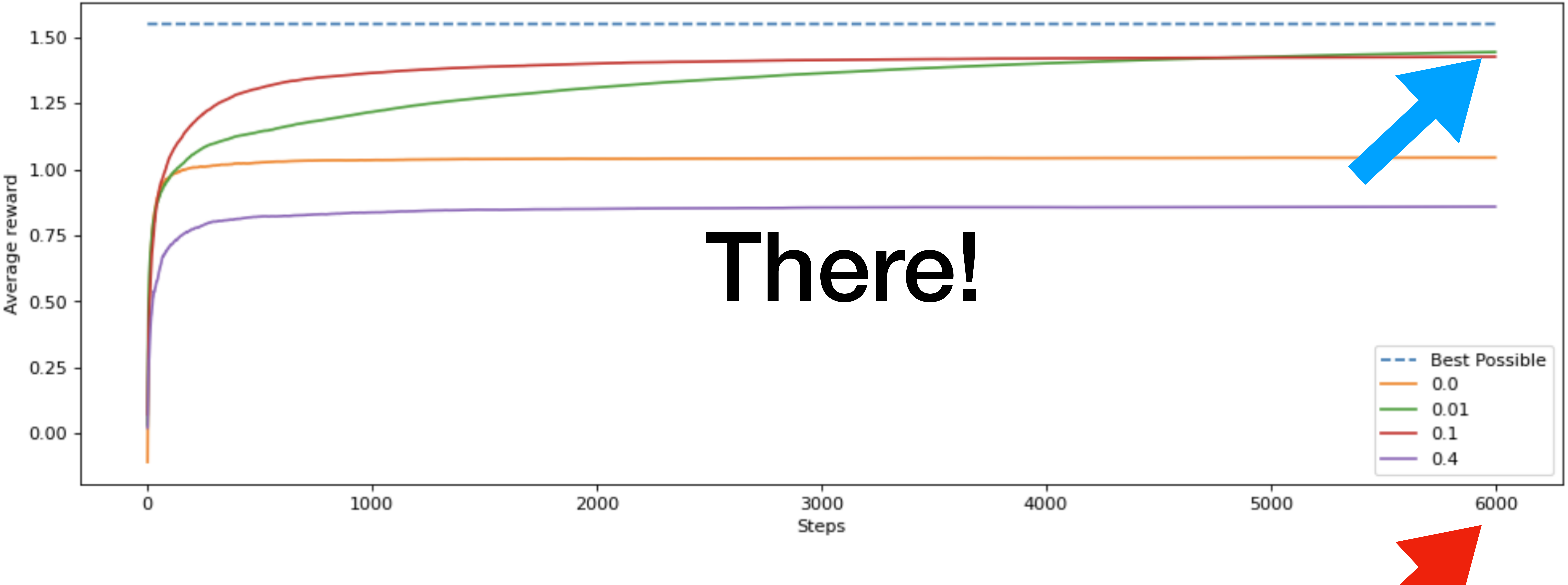
Will small epsilon do better?



Let the data guide you

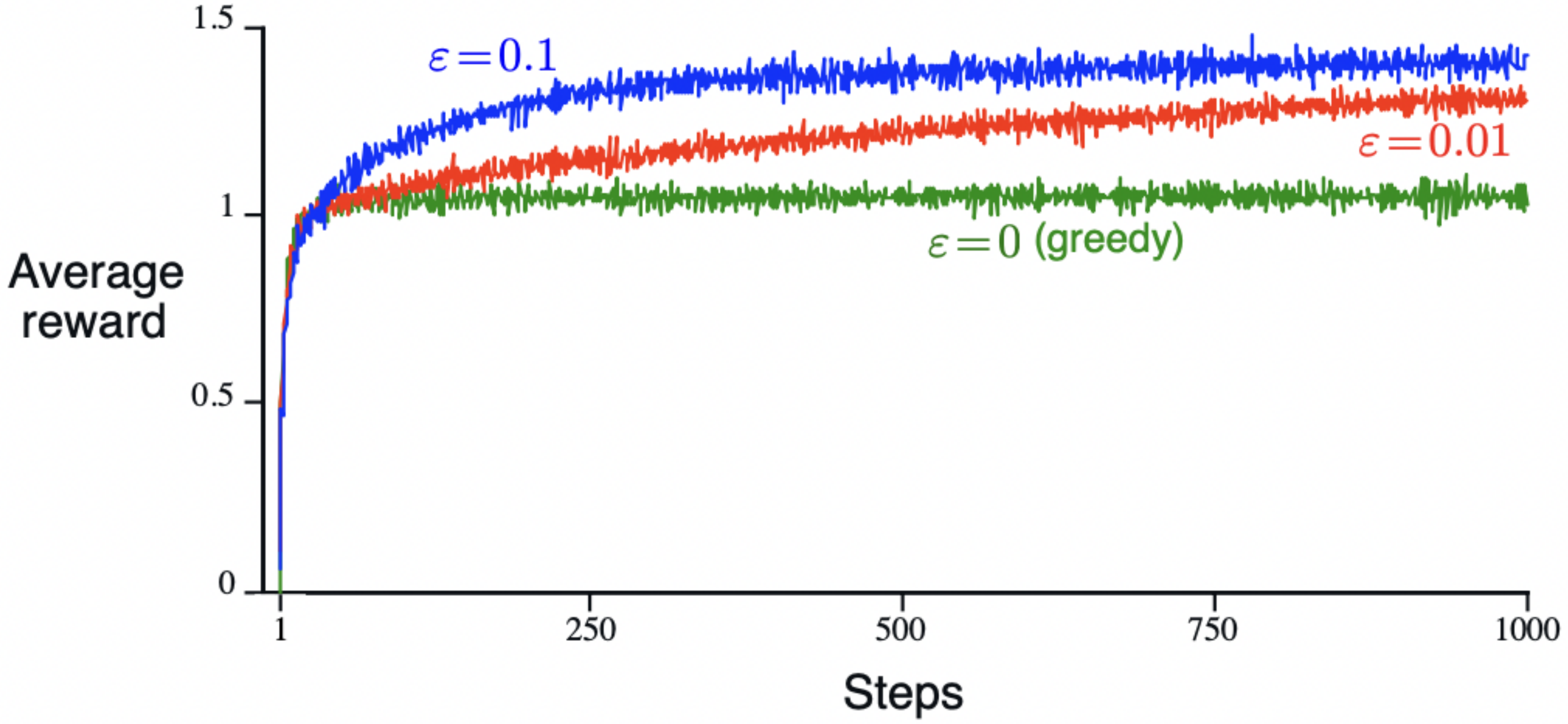


Let the data guide you

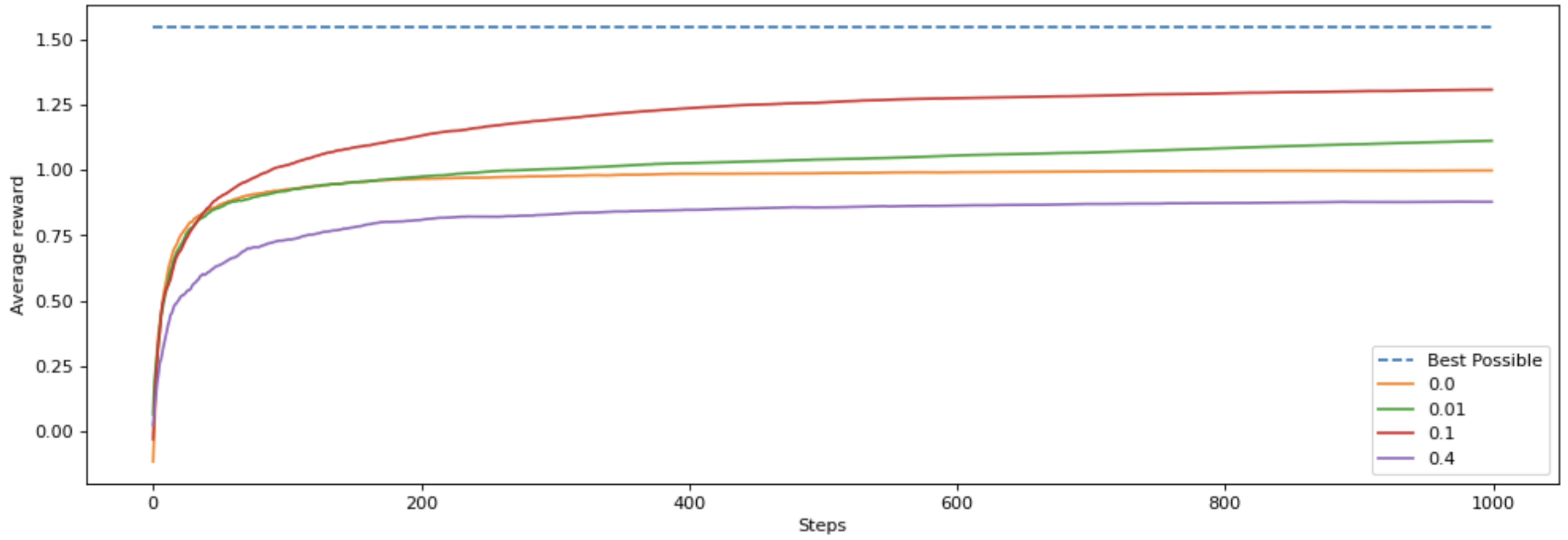


There!

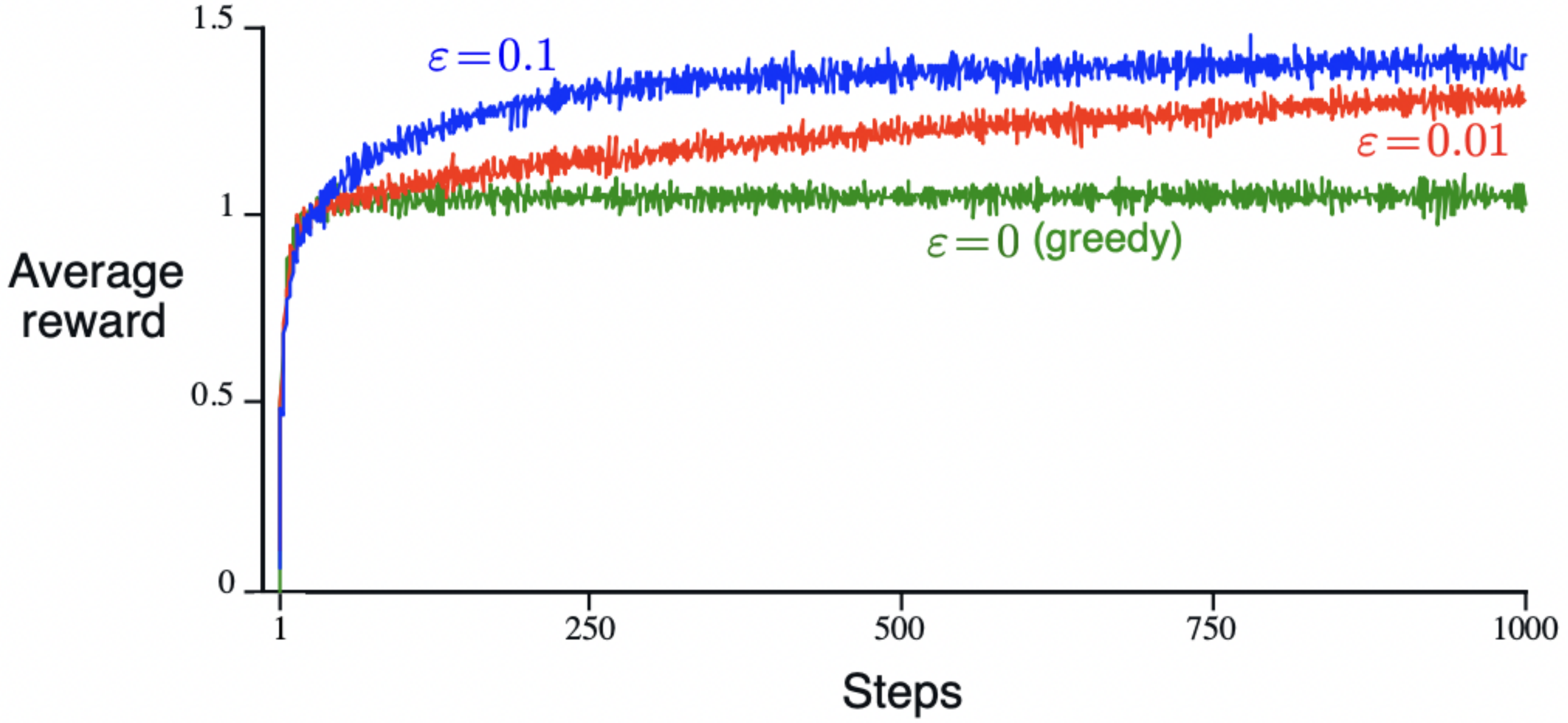
You might have noticed, the textbook curves are not as smooth!



VS

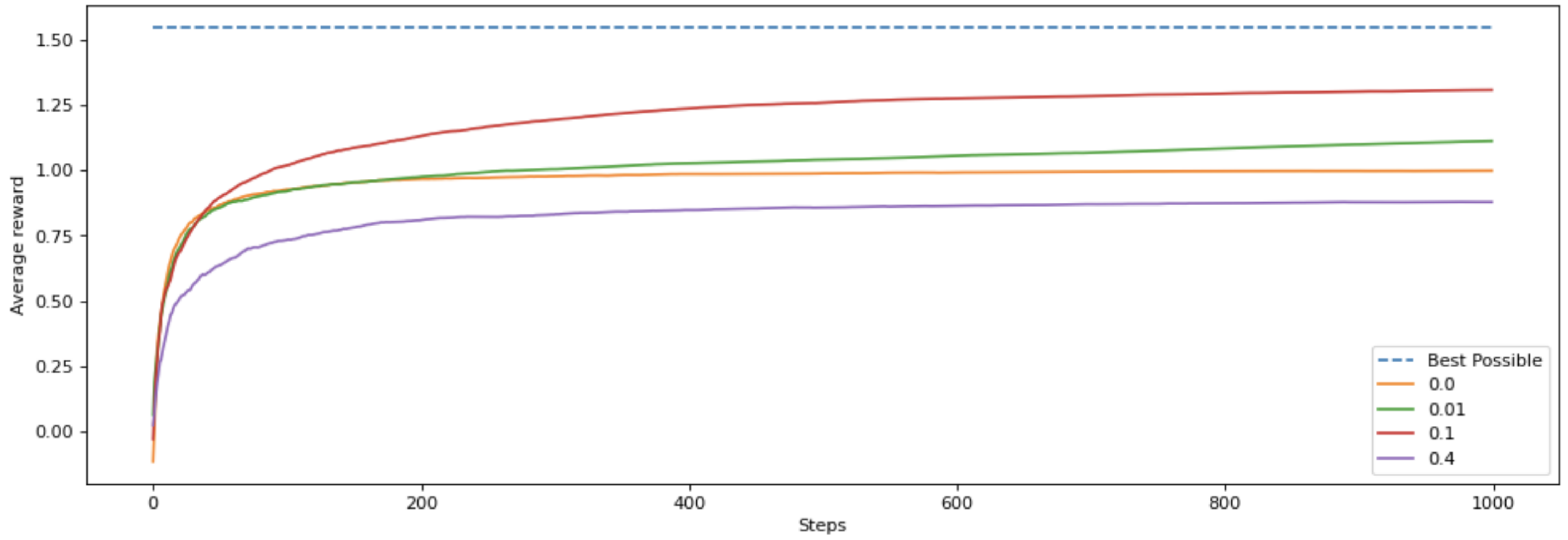


The textbook is only averaging overruns!

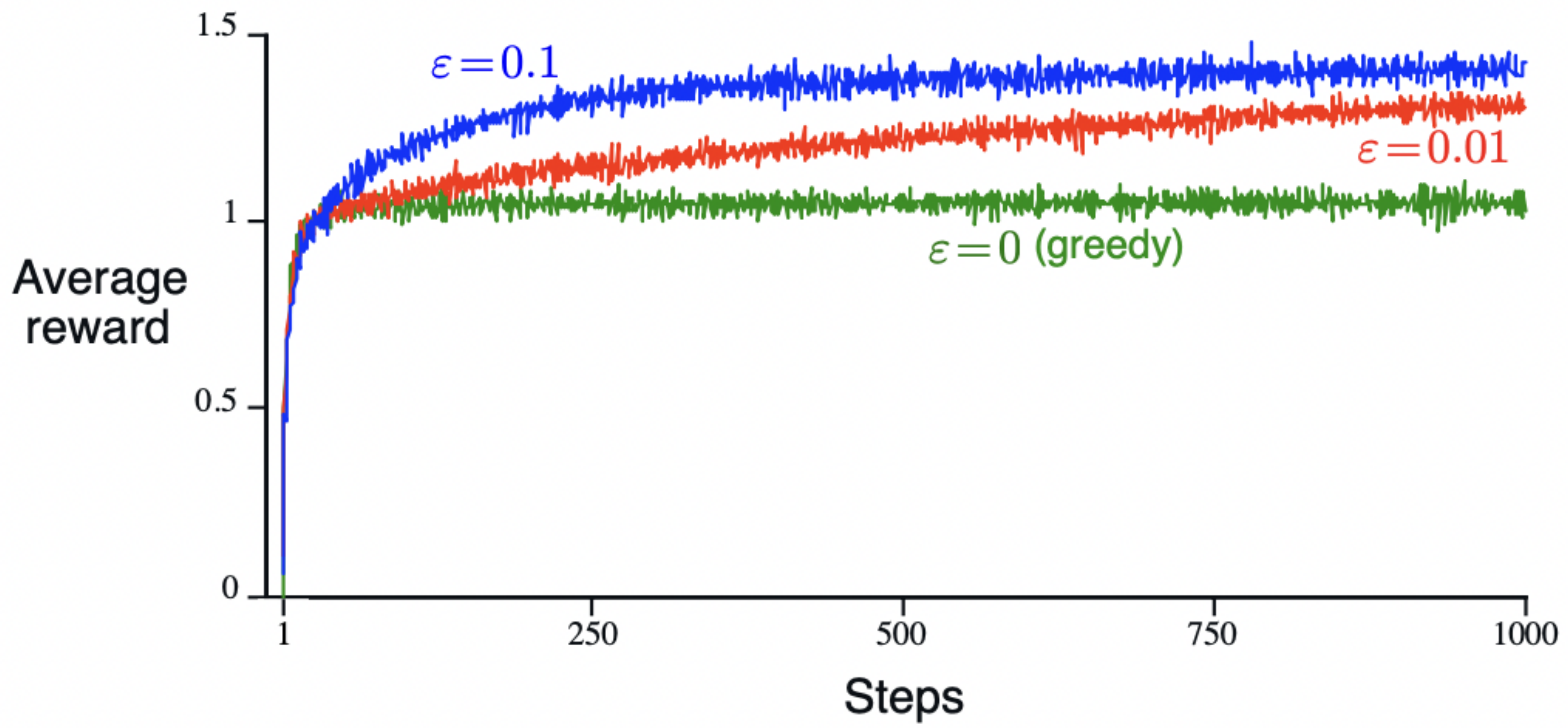


VS

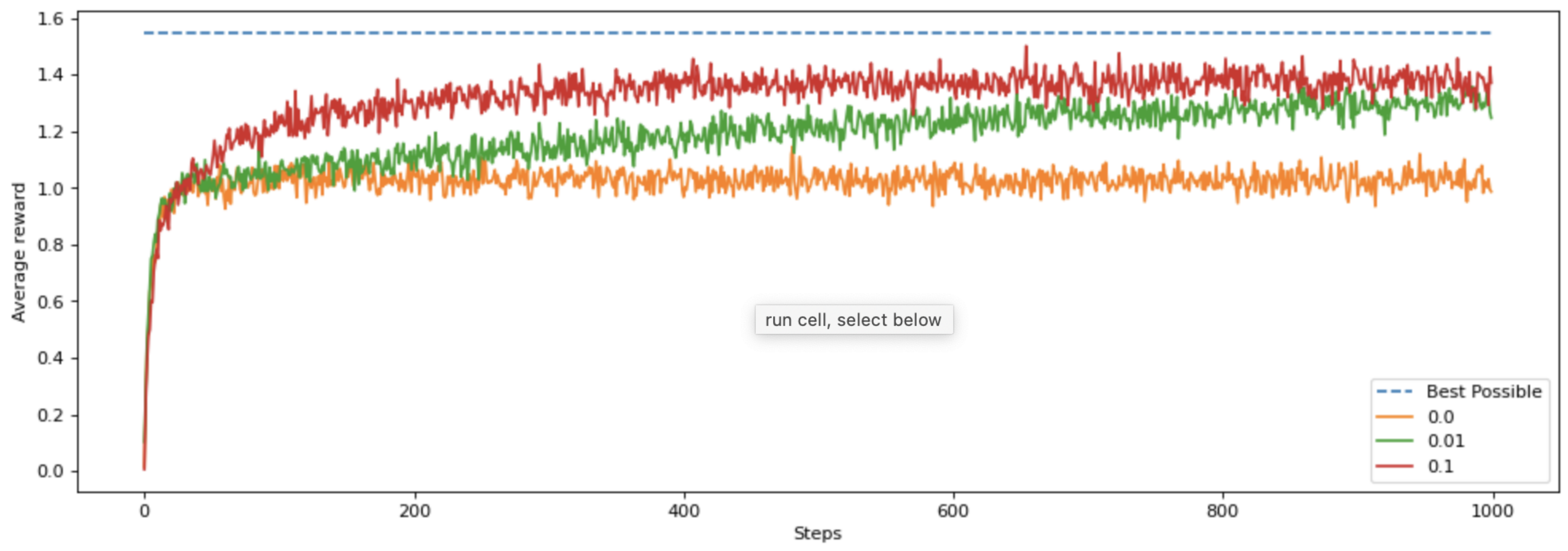
The notebook averaging the reward within the run first
&
using less runs (because compute :( )



Easy to fix



VS



Let’s review the quiz

- https://www.coursera.org/learn/fundamentals-of-reinforcement-learning/quiz/lMCZf/sequential-decision-making/attempt

Worksheet question

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1								
2								
3								
4								
5								

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $\boxed{R_1 = -1}$, $A_2 = 2$, $\boxed{R_2 = 1}$, $A_3 = 2$, $\boxed{R_3 = -2}$, $A_4 = 2$, $\boxed{R_4 = 2}$, $A_5 = 3$, $\boxed{R_5 = 0}$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1								
2								
3								
4								
5								

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $\boxed{R_1 = -1}$, $A_2 = 2$, $\boxed{R_2 = 1}$, $A_3 = 2$, $\boxed{R_3 = -2}$, $A_4 = 2$, $\boxed{R_4 = 2}$, $A_5 = 3$, $\boxed{R_5 = 0}$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1								-1
2								1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $\boxed{Q_1(a) = 0,}$ for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1								-1
2								1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $\boxed{Q_1(a) = 0,}$ for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0				-1
2								1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $\boxed{A_1 = 1,}$ $R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1		-1
2								1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $\boxed{A_1 = 1,}\ R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1	Maybe	-1
2								1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1	Maybe	-1
2								1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1	Maybe	-1
2	-1	0	0	0				1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1	Maybe	-1
2	-1	0	0	0	{2,3,4}			1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1	Maybe	-1
2	-1	0	0	0	{2,3,4}	2	Maybe	1
3								-2
4								2
5								0

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1	Maybe	-1
2	-1	0	0	0	{2,3,4}	2	Maybe	1
3	?	?	?	?				-2
4								2
5								0

Worksheet question

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

What is the definition of expected value?

What is the random variable & what are the possible outcomes?

Worksheet question

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

$$\mathbb{E}[X] \doteq \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$

Worksheet question

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

$$\mathbb{E}[X] \doteq \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$

Two outcomes: win or not

Worksheet question

1. 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.

(a) Model this as a K-armed bandit problem: define the action set.

Can you formally define q*?

(b) Is the reward a deterministic or stochastic function of your action?

(c) 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.

(d) Decide on which coin to flip next! Assume it's an exploit step.