Multi Agent Systems

Homework Assignment 4

Dirk Hoekstra, Luisa Ebner, Philipp Lintl

October 10, 2018

4 Coalitional Games and the Shapley Value

4.1 Building a new runway

4.1.1 Does that seem to be a fair division? How does that relate to the Shapley value?

This does seem like a fair division. As the first $\frac{1}{3}$ is used by all 3 companies each should pay $\frac{1}{3}$ of that cost. The next $\frac{1}{3}$ is used by 2 companies so each should pay $\frac{1}{2}$ of the cost. The final $\frac{1}{3}$ is used by 1 company so it should pay the full cost.

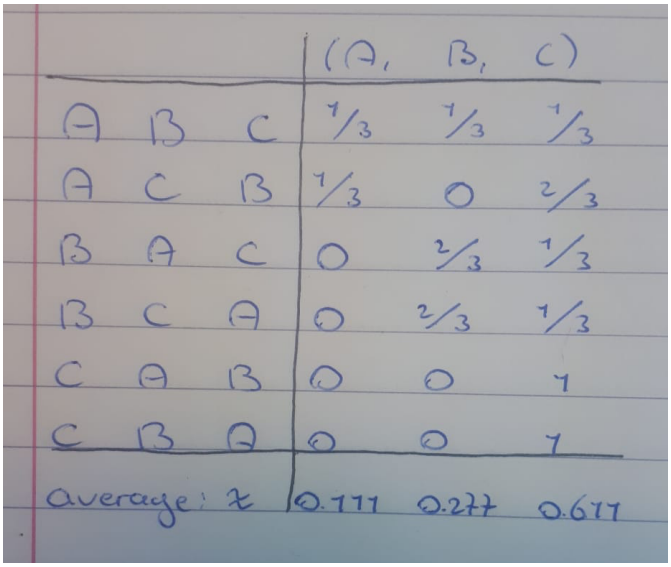
If we calculate this the percentage of the total cost that each company has to pay is as follows:

$$A = \frac{1}{3} * \frac{1}{3} \approx 0.111 \tag{1}$$

$$B = \frac{1}{3} * \frac{1}{3} + \frac{1}{3} * \frac{1}{2} \approx 0.278 \tag{2}$$

$$C = \frac{1}{3} * \frac{1}{3} + \frac{1}{3} * \frac{1}{2} + \frac{1}{3} * 1 \approx 0.611 \tag{3}$$

We can also calculate the Shapley value. See Figure 1.



			(A,	B,	C)
A	B	C	1/3	1/3	1/3
A	C	B	1/3	0	2/3
B	A	C	0	2/3	1/3
B	C	A	0	2/3	1/3
C	A	B	0	0	1
C	B	A	0	0	1
average:			0.111	0.277	0.611

Figure 1: Shapley Table

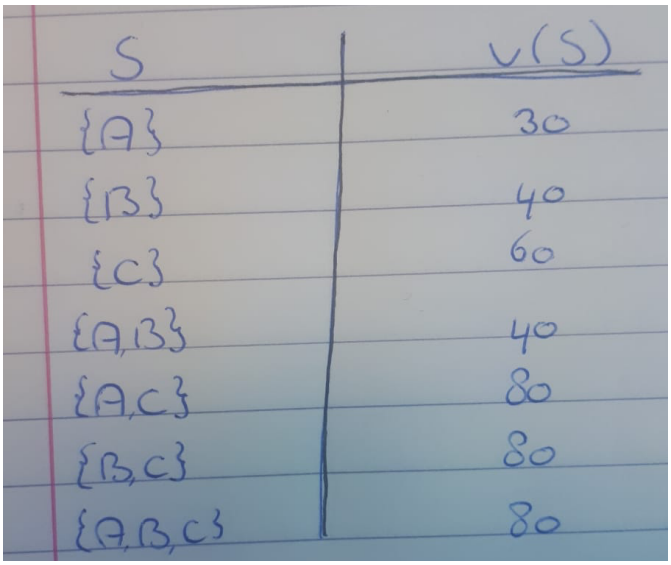
As you can see the calculating the Shapley value yields the same result. This means that A pays 1 million, B pays 2.5 million and C pays 5.5 million €.

4.2 Sharing a Taxi

4.2.1 What do you think would be a fair way to divide the taxi fare among them? Explain.

The simple way would be that each player pays $\frac{1}{3}$ of the cost, however this would be unfair as Charlie will pay €26.67 which is €33.33 less than if he would ride by himself. Whilst Alice only pays €3.33 less than if she where to ride by herself.

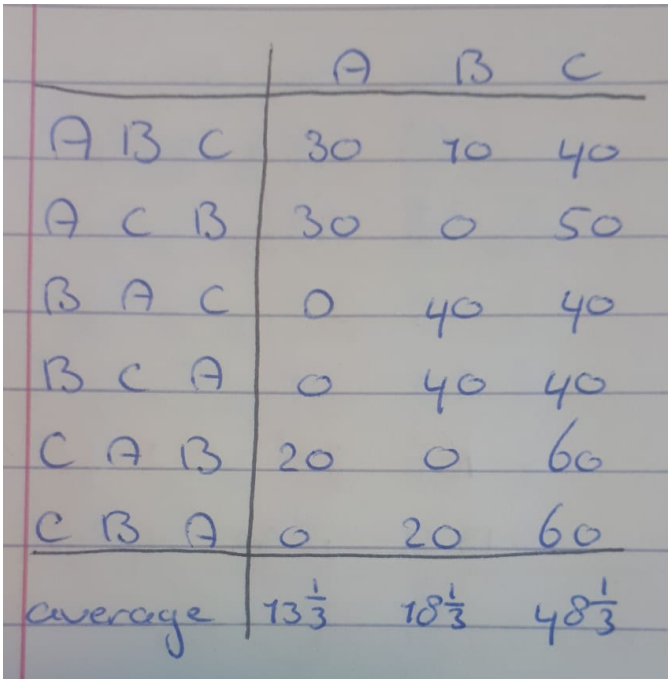
We should instead use the Shapley value to determine what each player should pay. We start with defining the value function in Figure 2.



S	v(S)
{A}	30
{B}	40
{C}	60
{A,B}	40
{A,C}	80
{B,C}	80
{A,B,C}	80

Figure 2: Value function

Next, using this value function we construct the table that determines the marginal contribution for each combination see Figure 3.



	A	B	C
A B C	30	10	40
A C B	30	0	50
B A C	0	40	40
B C A	0	40	40
C A B	20	0	60
C B A	0	20	60
average	$13\frac{1}{3}$	$18\frac{1}{3}$	$48\frac{1}{3}$

Figure 3: Marginal contributions

As you can see our conclusion is that Alice should pay €13.33, Bob should pay €18.33 and Charlie should pay €48.33

5 Exploitation versus Exploration

5.1 Kullback-Leibler divergence

5.1.1

Given are 2 continuous, one-dimensional probability densities f and g.

The Kullbach Leibler divergence is defined as $\int f(x)\, log\, \frac{f(x)}{g(x)} dx$

Let $f \sim N(\mu, \sigma^2)$ and $g \sim N(v, \tau^2)$.

5.1.1. Task: *Express $KL(f||g)$ as a function of mean and variance of f and g.*

$$
\begin{aligned}
&\int f(x)\, log\, \frac{f(x)}{g(x)} dx \\
=&\int f(x)\, log\, \frac{\frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\tau^2}} exp\left(\frac{-(x-v)^2}{2\tau^2}\right)} dx \\
=&\int f(x)\, log\, \left(\sqrt{\frac{\tau^2}{\sigma^2}}\, exp\, \left(\frac{-(x-\mu)^2}{2\sigma^2} + \frac{-(x-v)^2}{2\tau^2}\right)\right) dx \\
=&\int f(x)\, log\, \left(\sqrt{\frac{\tau^2}{\sigma^2}} dx + \int f(x)(\frac{-(x-\mu)^2}{2\sigma^2} + \frac{-(x-v)^2}{2\tau^2}\right) dx \\
=&\frac{1}{2}\, log\, \left(\frac{\tau^2}{\sigma^2}\right) + \frac{1}{2\sigma^2}\left(-\int (x-\mu)^2 f(x)dx) + \frac{1}{2\tau^2}(-\int (x-v)^2 f(x)dx\right) \\
=&\frac{1}{2}\, log\, \left(\frac{\tau^2}{\sigma^2}\right) - \frac{\sigma^2}{2\sigma^2} + \frac{1}{2\tau^2}\int (x-\mu+\mu-v)^2 f(x)dx \quad * \\
=&\frac{1}{2}\, log\, \left(\frac{\tau^2}{\sigma^2}\right) - \frac{1}{2} + \frac{1}{2\tau^2}\left(\int (x-\mu)^2 f(x)dx + (\mu-v)^2 \int f(x)dx + 2(\mu-v)\int (x-\mu) f(x)dx\right) \quad ** \\
=&\frac{1}{2}\, log\, \left(\frac{\tau^2}{\sigma^2}\right) - \frac{1}{2} + \frac{1}{2\tau^2}\left[\sigma^2 + (\mu-v)^2 + 0\right]
\end{aligned}
$$

$$
=\frac{1}{2}\left(log\left(\frac{\tau^2}{\sigma^2}\right) - 1 + \frac{\sigma^2 + (\mu-v)^2}{\tau^2}\right)
$$

$*$

- $\int (X-\mu)^2 f(x)dx = E\left((X-\mu)^2\right) = Var(X) = \sigma^2$

**

- $E\left((X-\mu)^2\right) = Var(X) = \sigma^2$
- $\int f(x)dx = 1$
- $E(X-\mu) = E(x) - \mu = \mu - \mu = 0$

5.1.2 Use the expression obtained above to argue that the KL divergence is always positive.

$$
\begin{aligned}
& \frac{1}{2}\left(log\left(\frac{\tau^2}{\sigma^2}\right) - 1 + \frac{\sigma^2 + (\mu - v)^2}{\tau^2}\right) \geq 0 \\
\leftrightarrow \quad & \left(log\left(\frac{\tau^2}{\sigma^2}\right) - 1 + \frac{\sigma^2 + (\mu - v)^2}{\tau^2}\right) \geq 0 \\
\leftrightarrow \quad & log\left(\frac{\tau^2}{\sigma^2}\right) - 1 + \frac{\tau^2}{\sigma^2} + \underbrace{\frac{(\mu - v)^2}{\tau^2}}_{\geq 0} \geq 0 \\
\leftrightarrow \quad & log\left(\frac{\tau^2}{\sigma^2}\right) + \frac{\tau^2}{\sigma^2} \geq 1
\end{aligned}
$$

Having tried several proof strategies and thus generally lacking in such, we decided to settle for a geometric argument.
As seen in Figure 4 the 3D plot of the left side of the last term never gets smaller than 1, which is exactly what we wanted to prove in the first place. For values close to zero the function reaches higher values and afterwards settles around 1 for larger τ^2 and σ^2 values.

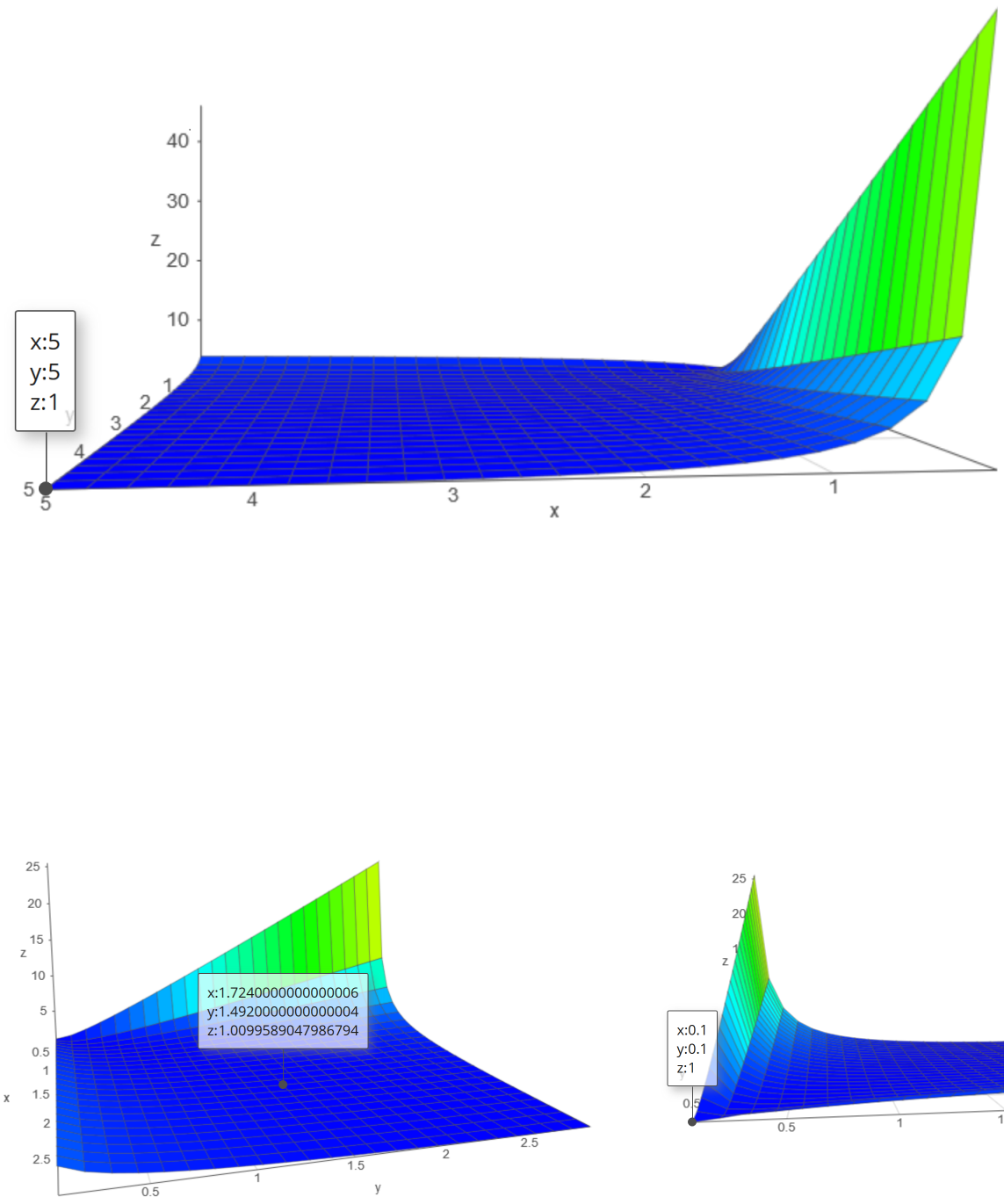


Figure 4: Different views on the function

5.1.3 Check your theoretical result in (1) by computing a sample-based estimate of the KL-divergence (Monte Carlo simulation). Pick an appropriate sample size.

Following the procedure for a Monte Carlo estimate of any KL divergence presented in Hershey and Olsen (2007) :

2. MONTE CARLO SAMPLING

The only method that really can estimate $D(f\|g)$ for large values of d with arbitrary accuracy is Monte Carlo simulation. The idea is to draw a sample x_i from the pdf f such that $E_f[\log f(x_i)/g(x_i)] = D(f\|g)$. Using n i.i.d. samples $\{x_i\}_{i=1}^n$ we have

$$D_{\mathrm{MC}}(f\|g) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i)/g(x_i) \to D(f\|g) \qquad (4)$$

Figure 5: Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models paper (DOI: 10.1109/ICASSP.2007.366913)

For a chosen example of $\mu_1 = 5, \sigma_1 = 2, \mu_2 = 7, \sigma_2 = 4$, f and g emerge as for n samples drawn from f:

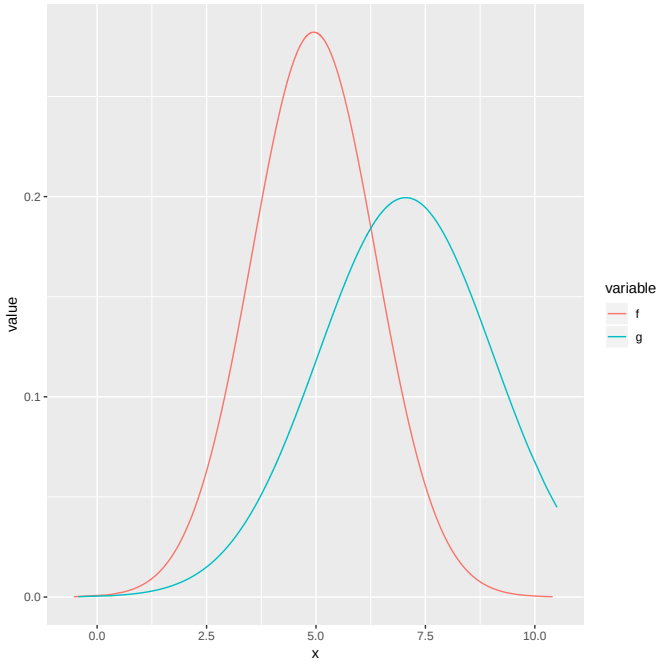


Figure 6: Densities of f and g

With those samples we conducted the Monte Carlo estimate, as well as the value obtained by plugging the parameter values into the form of above. We present both for several n values:

n	MC-KL	analytical-KL
10	0.8590996	0.5965736
100	0.7625015	0.5965736
1000	0.5770293	0.596573
10000	0.5924051	0.5965736
100000	0.5934350	0.5965736
1000000	0.5963060	0.5965736

Table 1: Analytical and Monte carlo estimate of KL for several sample sizes.

As seen in table 1, our obtained form of 5.1.1 gets very close to the Monte carlo estimate of KL divergence and thus can be supposed as correct.

5.2 The intuition behind Lai-Robbins lower bound for expected total regret

5.2.1 Why is it not optimal to stop sampling one of the arms after a certain time? Put differently, each arm should be sampled an infnite number of times (as t $\rightarrow$ inf) Why?

The k-armed bandit problem is a classical reinforcement learning task or problem as the agent is not told what to do in order to achieve his goals, but must discover the actions that yield his maximal reward by trying multiple times. As such, the task incorporates the conflicting needs of exploration and exploitation. On the one hand, the agent wants to explore actions that he has not tried before or rather minimize the uncertainty of the expected/mean reward coming with a certain action. On the other hand he wants to exploit what was already experienced as effective in producing reward. However, even if the first action turns out to have a higher expected reward after a finite number of time steps, it is not optimal to stop sampling from action two after this finite number of steps and only pull action one forever. As every action is given a probability distribution f_a, various rewards can be received with different probabilities. The rewards come along with uncertainty. The agent undertakes a stochastic task, in which each action must be tried many times to receive a reliable estimate of the expected gain. According to the Law of Large numbers, only after

$n \to \infty$, does the sample mean $Q_t(a)$ equal the actual mean reward $q(a)$. Before trying infinitely many times, Hoeffdings inequality claims that:

$$P(q(a) > Q_t(a) + U_t(a)) \leq e^{-2N_t(a)U_t(a)^2} = p(t)$$

The goal of ongoing exploration is a continuous minimization of the uncertainty along with the sample mean of $Q_t(a = 2)$. Before $N_t(a = 2) \to \infty$, there is a exceedance probability $p(t) > 0$ for the actual q_2 to be however much $(U_t(a = 2))$ larger than the current sample mean. In order to avoid getting stuck in suboptimal action taking of q_1, the agent has to keep on exploring from time to time also q_2. Only after both actions are played infinitely many times, does he know the actual mean reward of both actions. Only by exploring "forever" can he achieve his goal of maximizing the overall reward on the long run. As to that, Lai & Robbins claim that the overall regret L_t has a lower bound in terms of the Kullbach Leibler divergence, indicating that the total regret increases logarithmically with the number of time steps. The agent cannot reduce the overall regret by playing only q_1 after a certain, finite number of steps. The regret will furtherly increase due to the uncertainty/stochastic character of the game.

5.2.2 Can you see why (for large values of t) the expected total regret L_t would have a lower bound that is proportional to log t.

We want to assure that

$$EN_t(a) \overset{!}{\to} \infty \quad as \quad t \to \infty$$
$$with\ EN_t(a) = \sum_{i=1}^{t} P(A_i = a)$$

for the second arm (a=2). Therefore we compare the following sample schemes:

$$P(A_k = 2) = \frac{1}{k^\alpha}, \qquad with\ \alpha = \{1, 2, 1 + \epsilon\}\ (\epsilon > 0) \tag{4}$$

Considerations:

$$EN_{t\to\infty}(a = 2) = \sum_{i=1}^{\infty} P(A_i = 2) = \sum_{i=1}^{\infty} \frac{1}{k^\alpha} \tag{5}$$

Case analysis:

(i) $\alpha = 1$

$$EN_{t\to\infty}(a=2) = \sum_{i=1}^{\infty} \frac{1}{k} = \infty \quad (harmonic\ series) \tag{6}$$

$\Rightarrow \frac{1}{k}$ guarantees that the expected number of plays a=2 diverges to infinity as the number of steps increases to infinity.

(ii) $\alpha = 2$

$$EN_{t\to\infty}(a=2) = \sum_{i=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} \quad (Leonhard\ Euler) \tag{7}$$

(iii) $\alpha = 1 + \epsilon$

$$EN_{t\to\infty}(a=2) = \sum_{i=1}^{\infty} \frac{1}{k^{1+\epsilon}} \leq \frac{1}{1-2^{1-(1+\epsilon}} < \infty \tag{8}$$

$\Rightarrow$ The general harmonic series $\sum_{k=1}^{\infty} \frac{1}{k^{\alpha}}$ converges for all $\alpha > 1$ and thereby especially $\alpha = 2$ and $\alpha = 1 + \epsilon$. Case (ii) and (iii) do not ensure the claim stated above. As such they can be disregarded from here on.

We explicitely look at $P(A_k = 2) = \frac{1}{k}$ to show that the expected total regret L_t has a lower bound that is proportional to log t this proportionality is claimed by Lai Robbins. We suggest a geometric argument to support our point.

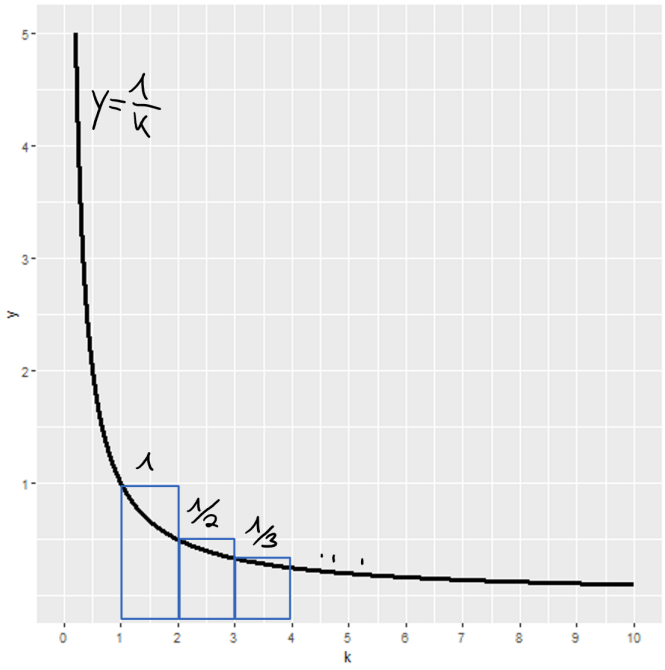


Figure 7

As shown the expected regret then emerges as

$$L_t = \sum_a \triangle_a EN_t(a) = \triangle_2 EN_t(2) \tag{9}$$

with $\triangle_2 = c$:

$$= c \cdot \sum_{k=1}^{t} P(A_k = 2) = c \cdot \sum_{k=1}^{t} \frac{1}{k} \tag{10}$$

As seen in Figure 7 this sum is larger than the appropriate integral:

$$c \cdot \sum_{k=1}^{t} \frac{1}{k} > c \cdot \int_1^{t+1} \frac{1}{k} dk > c \cdot \int_1^{t} \frac{1}{k} dk = c \cdot ln(t) \tag{11}$$

5.3 UCB vs ϵ-greedy

5.3.1 Write a programme to experiment with the exploration/exploitation for the k-bandit problem (e.g.take 5≤k≤ 20). Assume that the arms generate normally distributed rewards. Produce graphs to compare the average reward (over time) for different strategies (greedy, greedy with optimistic initialization, UCB).

UCB strategy

When using the UCB strategy the player will first choose an arm that hasn't been pulled before. Then the program will choose the arm with the highest possible mean: "optimism in the face of uncertainty". This produces the following graph. You can see a nice dip in the beginning where the algorithm is still figuring out wich arm is the best, but once it has decided it will keep playing that arm getting close to the real max mean.

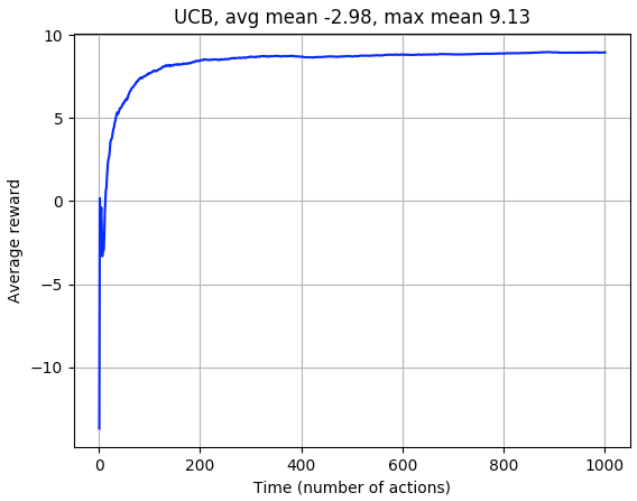


Figure 8: UCB graph

ϵ Greedy

Greedy will choose the arm with the highest observed mean with probability $(1 - \epsilon)$ and a random arm otherwise (with probability ϵ. We see in Figure 9 that the $\epsilon = 0.1$ player initially choose the right arm to play and kept playing that arm most of the time (because the probability of exploring other arms is only 10%). However we not that the average reward does not come as close to the max mean as in the UCB strategy.

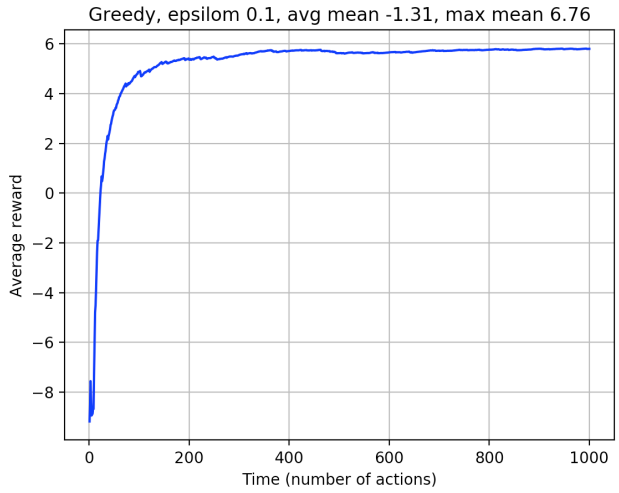


Figure 9: Greedy with $\epsilon = 0.1$

We see in Figure **??** that with $\epsilon = 0.5$ that the player almost immediately found a good arm to play, however in the long term it keeps exploring even if it already found the best arm this explains the dip and why the line is not close to the maximum mean.

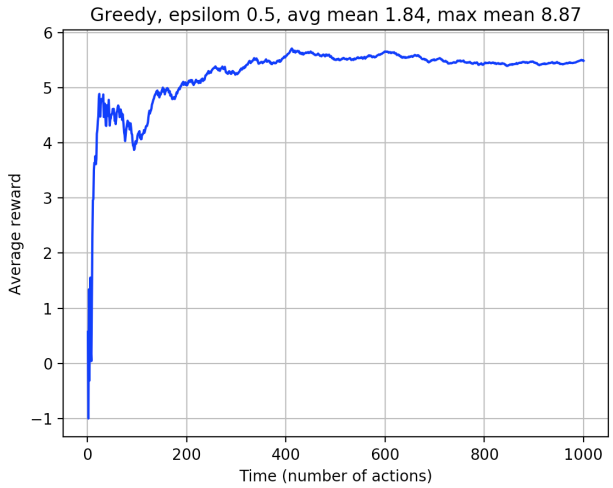


Figure 10: Greedy with $\epsilon = 0.5$

We see in Figure 11 with $\epsilon = 0.9$ that it is almost totally random. The player will play the best arm only 10% of the time. This leads to an average that is barely higher than the real average mean.

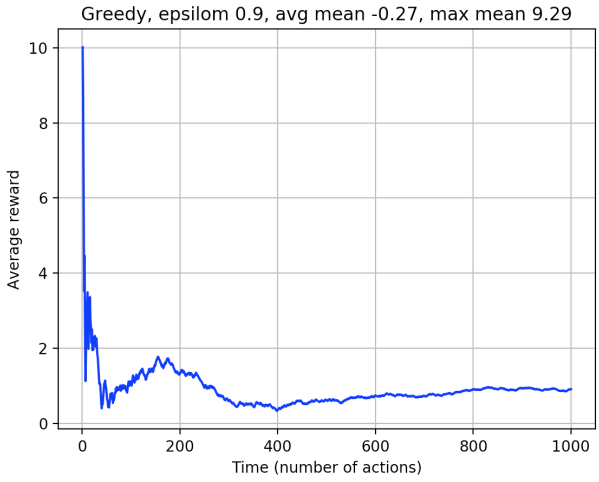


Figure 11: Greedy with $\epsilon = 0.9$

Greedy with optimistic initialization

With the optimistic greedy we set the observed mean to $maxmean + 10$. This means that in the beginning the bandit will play all arms, and the observed mean will drop for each arm. Then, it will find an arm that does not drop and will keep choosing that, which results in the nice graph that is shown in Figure 12

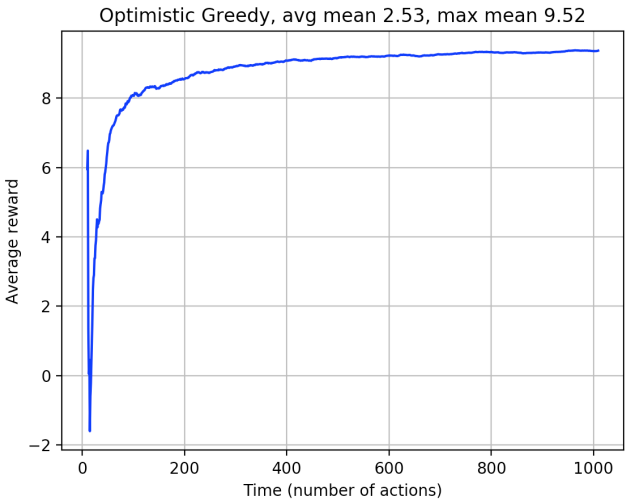


Figure 12: Optimistic Greedy

Conclusion

Greedy with the right ϵ value produces better results in the short run. UCB gets very close to the actual max mean in the long run. Optimistic greedy is very similar to the UCB and produces similar results in the long run.