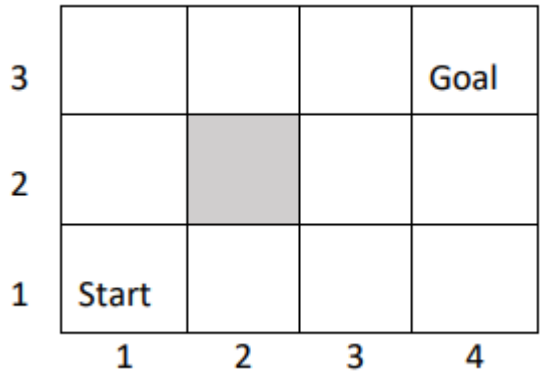
INFORME PRÀCTICA 3

INTEL·LIGÈNCIA ARTIFICIAL

Pau Segura
David Vllajosana

12 de desembre de 2023

1. We will start by tackling the simple problem seen in class of finding a path from start to goal in the following scenario:



Implement the Q-learning algorithm to find the optimal path considering a reward of -1 everywhere except for the goal, with reward 100.

i) Print the first, two intermediate and the final Q-table. What sequence of actions do you obtain?

PRIMERA Q-TABLE I POLITICA

[-0.2, 0, -inf, -inf]
[-0.2, -inf, 0, -inf]
[-inf, -0.2, 0, -inf]
[-inf, 0, -inf, 0]
[-inf, -0.2, -inf, 0]
[0, 0, -inf, 0]
[0, 0, 0, -inf]
[-inf, 20.0, 0, 0]
[0, -inf, -inf, 0]
[0, -inf, 0, 0]
[0, 0, 0, 0]

['↓', '←', '→', '↑']
['↓', 'o', '↑', '↑']
['→', '→', '↑', '↑']

SEGONA Q-TABLE I POLITICA

[57.6190313445927, 62.17099999999989, -inf, -inf]
[68.19603506983192, -inf, 26.33692171216058, -inf]
[-inf, 78.53643201466653, 11.24722372362896, -inf]
[-inf, 70.1899999999999, -inf, -0.771660416]

[-inf, 88.92976931363623, -inf, 0]
[79.09999999999992, 79.01733366662371, -inf, -0.28784000000000004]
[88.99999999999994, 86.85845876273467, 69.36657902087684, -inf]
[-inf, 99.99999999999997, 78.78078110046174, 0]
[88.99604008357991, -inf, -inf, 0]
[99.9999215362283, -inf, 62.52835811348432, 0]
[0, 0, 0, 0]

['→', '→', '→', '↑']
['↑', 'o', '↑', '↑']
['→', '→', '↑', '↑']

TERCERA Q - TABLE I POLITICA

[62.05737509359908, 62.17099999999989, -inf, -inf]
[70.16135303667576, -inf, 40.30200711662617, -inf]
[-inf, 79.09452788825821, 21.425805715614413, -inf]
[-inf, 70.1899999999999, -inf, -0.771660416]
[-inf, 88.99958543120079, -inf, 0]
[79.09999999999992, 79.09993459793057, -inf, -0.28784000000000004]
[88.99999999999994, 88.99585757191441, 70.18688920349344, -inf]
[-inf, 99.99999999999997, 79.09903522007555, 0]
[88.99999913845429, -inf, -inf, 0]
[99.99999999990379, -inf, 73.66971619101255, 0]
[0, 0, 0, 0]

['→', '→', '→', '↑']
['↑', 'o', '↑', '↑']
['→', '→', '↑', '↑']

Q-TABLE FINAL I LA SEVA POLITICA
[62.1707223259512, 62.17099999999989, -inf, -inf]
[70.18995495397517, -inf, 45.57668855464072, -inf]
[-inf, 79.09999329966864, 41.307787397005264, -inf]
[-inf, 70.1899999999999, -inf, -0.771660416]
[-inf, 88.99999967154473, -inf, 0]
[79.09999999999992, 79.09999974508632, -inf, -0.28784000000000004]
[88.99999999999994, 88.99981781408384, 70.18998530967878, -inf]
[-inf, 99.99999999999997, 79.09999880566026, 0]
[88.99999999931704, -inf, -inf, 0]
[99.99999999999997, -inf, 77.96118686538694, 0]
[0, 0, 0, 0]

['→', '→', '→', '↑']
['↑', 'o', '↑', '↑']
['→', '→', '↑', '↑']

Hem decidit, mostrar com a taules intermèdies les taules situades a ⅓ de la llista de Q-tables, i als ⅔.
A mesura que un agent dins d'un marc d'aprenentatge per reforç completa més episodis, va acumulant una millor experiència, és a dir, un millor aprenentatge. Aquesta experiència és fonamental per permetre que l'algorisme pugui afinar de manera més precisa els valors en la taula Q.

En el primer episodi, la presa de decisions de l'agent és arbitrària, ja que el valor màxim de la Q-table és el mateix en totes les accions de cada estat.

ii) After trying for a bit, what is your parameter choice for alpha, gamma and epsilon? Why?

Alpha = 0.2, Gamma = 0.9, Epsilon = 0.15.
Hem triat aquests paràmetres, ja que hem vist que amb aquests, l'agent aconsegueix un equilibri òptim entre aprenentatge i estabilitat. Amb aquests valors, no hi overfitting de valors q ni una gran generació de moviemnts aleatoris.

iii) How do you judge convergence of the algorithm? How long does it take to converge?

Jutgem en funció de l'estabilitat dels q-values. Hem definit un llindar de convergència(**0.0001)** i un mínim de coincidències. Si es dona el cas que la resta en valor absolut de la mitja dels valors q de l'episodi anterior amb la mitja dels valors de l'episodi actual és menor al llindar durant 100 iteracions seguides, vol dir què convergeix.
Calculant la mitja de 5 execucions de l'algorisme amb 4000 episodis., trobem que ha convergit sobre l'episodi **1832.2.**

b) Try implementing the more accurate reward given by:

3	-3	-2	-1	100
2	-4		-2	-1
1	-5	-4	-3	-2
	1	2	3	4

i.i) Answer the questions of the previous section for this case.

i.i)
PRIMERA Q-TABLE I POLITÍCA
[-0.4, 0, -inf, -inf]
[-0.30000000000000004, -inf, 0, -inf]
[-inf, -0.2, 0, -inf]
[-inf, 0, -inf, 0]
[-inf, -0.1, -inf, 0]
[0, 0, -inf, 0]
[0, 0, 0, -inf]
[-inf, 10.0, 0, 0]
[0, -inf, -inf, 0]
[0, -inf, 0, 0]
[0, 0, 0, 0]

['↓', '←', '→', '↑']
['↓', 'o', '↑', '↑']
['→', '→', '↑', '↑']

SEGONA Q-TABLE I POLÍTICA

[56.56099999999978, 56.55826898925741, -inf, -inf]
[67.2899999999998, -inf, 45.90470448020296, -inf]
[-inf, 78.09999999999982, 56.560766530126635, -inf]
[-inf, 67.2896073255096, -inf, -2.3352167170022002]
[-inf, 88.99999999999989, -inf, -0.33420000000000005]
[78.09997817552168, 41.29059840851791, -inf, -0.5027939536]
[88.99999999979609, 74.81136451321993, 54.59547416873384, -inf]
[-inf, 99.99999999999994, 78.0999344934502, 0]
[76.86099821343817, -inf, -inf, 0]
[99.21448327887208, -inf, 19.297874724432152, 0]
[0, 0, 0, 0]

['→', '→', '→', '↑']
['↑', 'o', '↑', '↑']
['↑', '→', '↑', '↑']

TERCERA Q-TABLE I POLÍTICA
[56.5609999999997, 56.5609999938541, -inf, -inf]
[67.289999999999, -inf, 45.90489999997813, -inf]
[-inf, 78.09999999999982, 56.56099999988666, -inf]
[-inf, 67.28999999957165, -inf, -2.3352167170022002]
[-inf, 88.99999999999989, -inf, -0.33420000000000005]
[78.09999999998378, 70.17524504464672, -inf, -0.5027939536]
[88.99999999999989, 84.43966300920673, 65.17291056853674, -inf]
[-inf, 99.99999999999994, 78.09999999999093, 0]
[87.5477947727581, -inf, -inf, 0]

[99.96670103634682, -inf, 29.63042847918542, 0]
[0, 0, 0, 0]

['→', '→', '→', '↑']
['↑', 'o', '↑', '↑']
['↑', '→', '↑', '↑']

QUARTA Q-TABLE I POLÍTICA

[56.56099999999978, 56.56099999999978, -inf, -inf]
[67.2899999999998, -inf, 45.90489999999977, -inf]
[-inf, 78.09999999999982, 56.56099999999978, -inf]
[-inf, 67.2899999999998, -inf, -2.3352167170022002]
[-inf, 88.99999999999989, -inf, -0.33420000000000005]
[78.09999999999982, 75.6311846042993, -inf, -0.5027939536]
[88.99999999999989, 88.44114305731415, 67.13801457573845, -inf]
[-inf, 99.99999999999994, 78.09999999999982, 0]
[88.66290100922296, -inf, -inf, 0]
[99.99897095698543, -inf, 34.39952973600615, 0]
[0, 0, 0, 0]

['→', '→', '→', '↑']
['↑', 'o', '↑', '↑']
['↑', '→', '↑', '↑']

Podem observar que com l'exercici anterior a la segona q-table ja ha trobat la política correcta.

i.ii) After trying for a bit, what is your parameter choice for alpha, gamma and epsilon? Why?

Alpha = 0.2, Gamma = 0.9, Epsilon = 0.15.
Hem triat aquests paràmetres, ja que hem vist que amb aquests, l'agent aconsegueix un equilibri òptim entre aprenentatge i estabilitat. Amb aquests valors, no hi overfitting de valors q ni una gran generació de moviments aleatoris.

i.iii) How do you judge convergence of the algorithm? How long does it take to converge?

Jutgem en funció de l'estabilitat dels q-values. Hem definit un llindar de convergència(**0.001)** i un mínim de coincidències. Si es dona el cas que la resta en valor absolut de la mitja dels valors q de l'episodi anterior amb la mitja dels valors de l'episodi actual és menor al llindar durant 100 iteracions seguides, vol dir què convergeix.
Fent la mitja feta també a l'exercici a, trobem que ha convergit en una mitja de **2675.4** episodis. Té sentit que el número sigui més gran, ja que si durant els

episodis fa una acció aleatòria cap a un estat amb recompensa inferior, la q-table es veurà modificada de manera més significativa i la diferència entre les q-tables dels episodis serà més gran que el llindar.
Trobem que el fet que sigui un algorisme que contempla l'estocàstica el resultat varia molt en funció de l'execució, i els episodis de convergència són diferents en funció de l'atzar, amb aquest llindar tan baix.

ii) What is the effect of the new reward function on performance?

L'efecte es veu en els episodis de convergència com hem explicat en l'exercici anterior. Si només valoressim el millor moviment que pot fer, trobaria el camí més ràpid que en una taula amb recompenses de -1, ja que les caselles més properes al goal tenen un valor més proper a 0. La lògica de l'algorisme faria avançar a l'agent sempre cap endavant, ja que trobaria que és la millor acció possible perquè la recompensa és major.

iii) How does this relate to the search algorithms studied in P1? Could you apply one of those in this case?

Si en el q-learning només féssim la millor opció seria semblant a l'A*, però ens quedarem sense el factor estocàstic que ens permet trobar un possible millor moviment.
La base del q-learning és l'exploració vs explotació i com no és un escenari determinista no podem aplicar cap algorisme vist a la P1.

c) The main novelty in RL algorithms with respect to the search algorithms in P1 is that they can be applied in stochastic environments, where the agent doesn't fully determine the outcome of its actions.

i) Drunken sailor.

ii)

1. **What is your parameter choice? Why?**
   Hem mantingut els mateixos valors, ja que són els que millor ens funcionen en termes d'aprenentatge i estabilitat

2. **Assuming the sailor is in a state that allows learning: how many drunken nights are necessary for them to master the perilous path to bed? Compare to the previous, deterministic scenario.**

   Com el factor de fer el pas correcte és del 99%, no trobem gaire diferència amb la quantitat de nits (episodis) que triga a aprendre el camí. El troba al

voltant dels 2000 episodis. Si el factor de fer el pas correcte disminueix creiem que trigaria més episodis a convergir.

3. **What is the optimal path found? If we watched the sailor try to follow it, would they always follow the same path?**

   ['→', '→', '→', '↑']
   ['↑', 'o', '↑', '↑']
   ['↑', '→', '↑', '↑']

   No, no el seguiria sempre, ja que comptem amb el factor estocàstic que el podria fer desviar a cada iteració.

4. **Could we apply one of the algorithms in P1 here? Why? Hint: think of the notions of deterministic vs random and of path vs policy.**

   Els algoritmes AStar, BFS i DFS són deterministes i estan dissenyats per a entorns on les conseqüències de les accions són previsibles i constants. Aquests algoritmes són efectius per trobar un camí específic d'un punt a un altre, però no s'adapten bé a entorns on les respostes a les accions poden ser imprevisibles o aleatòries, com és el cas en el Q-learning amb un factor estocàstic.

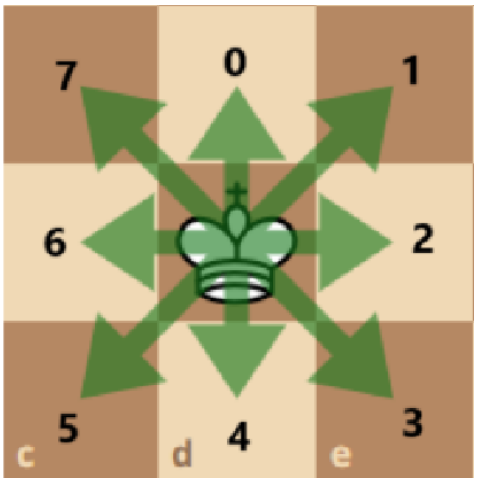
2. Now, let's move back to the chess scenario, namely the first board configuration of P1. Remember that we have the black king, the white king and one white rook, and that only whites move. Remember to provide the first, two intermediate and the final Q-table in every case.

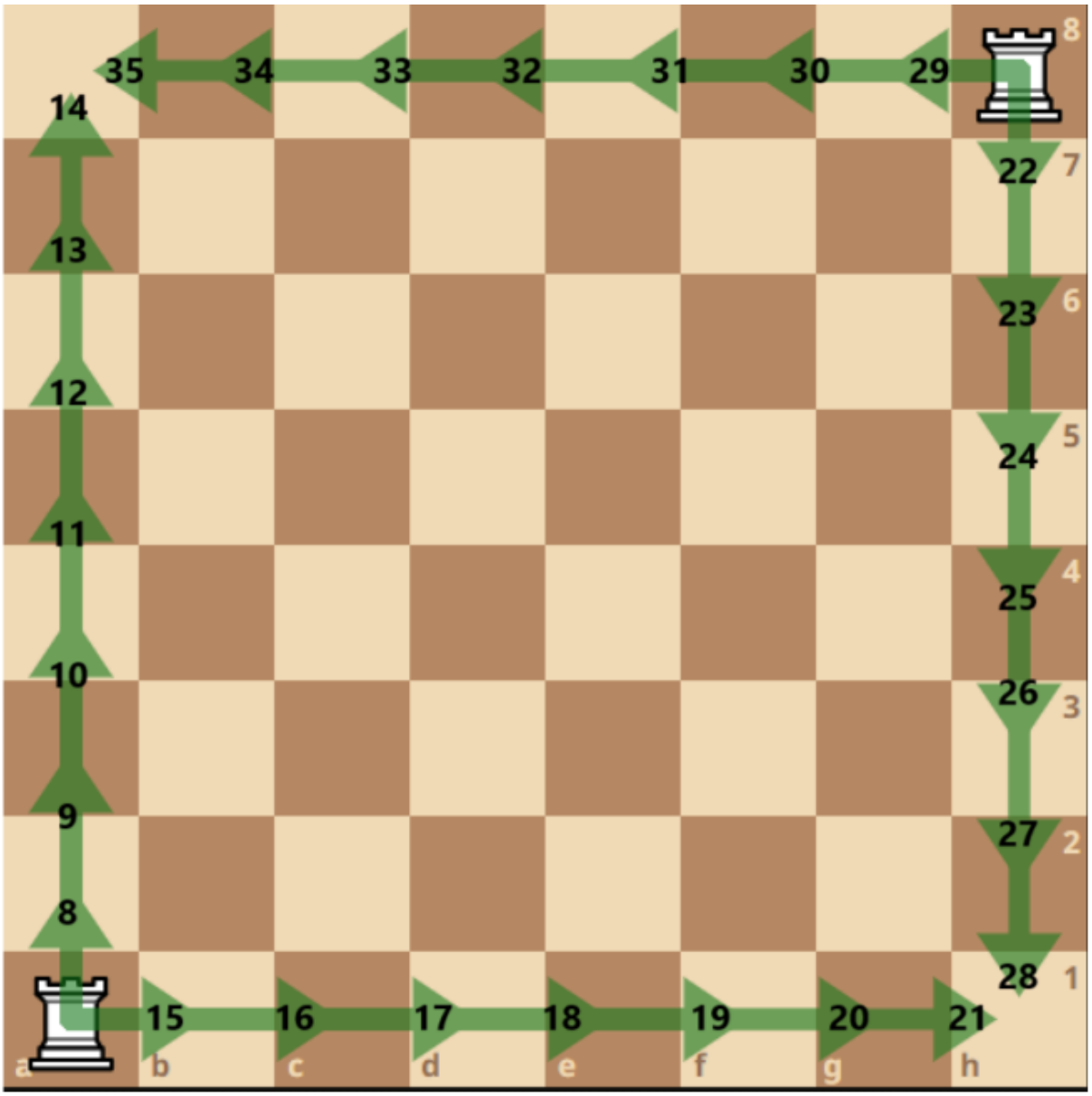
Les q-tables estan adjuntades al final del document.

a.Adapt your Q-learning implementation to find the optimal path to a check mate considering a reward of -1 everywhere except for the goal (check mate for the whites), with reward 100.

i) What sequence of actions do you obtain?

Glossari accions(els mmoviments ilegals estan a la QTable com a valor -infinit):





La seqüència que obtenim és:
estado: [[7, 0, 2], [7, 4, 6]]
ACCIO: 0
estado: [[0, 4, 12], [7, 0, 2], [6, 4, 6]]
ACCIO: 14
estado: [[0, 4, 12], [0, 0, 2], [6, 4, 6]]
ACCIO: 1
estado: [[0, 4, 12], [0, 0, 2], [5, 5, 6]]
ACCIO: 1
estado: [[0, 4, 12], [0, 0, 2], [4, 6, 6]]
ACCIO: 16
estado: [[0, 4, 12], [0, 2, 2], [4, 6, 6]]
ACCIO: 7
estado: [[0, 4, 12], [0, 2, 2], [3, 5, 6]]
ACCIO: 7

ii) After trying for a bit, what is your parameter choice for alpha, gamma and epsilon? Why?

En aquest cas, el learning rate és 0,4, la discount 0,8 i l'èpsilon és 0,5. En aquest exercici hem determinat una èpsilon inicial bastant alt i l'anem disminuint a cada vegada que fem un nou episodi. Ho fem així, ja que creiem que a les primeres iteracions és més important explorar, ja que no coneix encara la majoria dels valors.

iii) How do you judge convergence of the algorithm? How long does it take to converge?

Jutgem en funció de l'estabilitat dels q-values. Hem definit un llindar de convergència(**0.000001)** i un mínim de coincidències. Si es dona el cas que la resta en valor absolut de la mitja dels valors q de l'episodi anterior amb la mitja dels valors de l'episodi actual és menor al llindar durant 100 iteracions seguides, vol dir què convergeix.

b. Try now with a more sensible reward function adapted from the heuristic used for the A* search:

i.) Answer the questions of the previous section for this case.

i.i) What sequence of actions do you obtain?

La seqüència que obtenim és:
estado: [[7, 0, 2], [7, 4, 6]]
ACCIO: 0
estado: [[0, 4, 12], [7, 0, 2], [6, 4, 6]]
ACCIO: 14
estado: [[0, 4, 12], [0, 0, 2], [6, 4, 6]]
ACCIO: 1
estado: [[0, 4, 12], [0, 0, 2], [5, 5, 6]]
ACCIO: 1
estado: [[0, 4, 12], [0, 0, 2], [4, 6, 6]]
ACCIO: 16
estado: [[0, 4, 12], [0, 2, 2], [4, 6, 6]]
ACCIO: 7
estado: [[0, 4, 12], [0, 2, 2], [3, 5, 6]]
ACCIO: 7

i.ii) After trying for a bit, what is your parameter choice for alpha, gamma and epsilon? Why?

Hem utilitzat la mateixa configuració que amb les recompenses de 100 i -1, ja que també ens hem adonat que amb una èpsilon major a l'inici troba abans el camí òptim.

ii.) What is the effect of the new reward function on performance?

Usant l'heurística, com més lluny sigui l'estat actual d'algun dels estats que fan escac i mac, els q-values eran més petits, ja que canviem el -1 de la recompensa pel valor de l'heurística, que com vem observar a la P1, són valors molt més petits.

L'algorisme triga més temps a calcular l'heurísitica que no pas la recompensa, així que l'execució és més lenta.

c.) Drunken sailor. On their way to bed, our drunken sailor sees a chessboard on a table, coincidentally configured as in the previous section. They have seen the captain play with the first mate and want to give it a try, but only have a rudimentary knowledge of the rules (they know how each piece moves and what is a check mate, but not that blacks move as well).

i.) Introduce stochasticity (= randomness) by enforcing that only a given percentage of the moves intended by the sailor are actually taken, the rest taken randomly from all other possibilities.

ii.) Use any reward you prefer:

1. **What is your parameter choice? Why?**

   Hem utilitzat la mateixa configuració que en el context sense Drunken Sailor, ja que també ens hem adonat que amb una èpsilon major a l'inici troba abans el camí òptim.

2. **Assuming our obsessive sailor is in a state that allows learning: how many games do they have to play before they are satisfied that they have found the best strategy and can go to bed? Compare to the previous, deterministic scenario.**

   El mariner ha trigat 850 partides a aprendre el camí òptim a l'escac i mat. Hem canviat el criteri de convergència i hem rebaixat els episodis consecutius convergents a 40.

3. **What is the optimal path found? If we watched the sailor try to follow it, would they always follow the same path?**

   estado: [[7, 0, 2], [7, 4, 6]]
   ACCIO: 14
   estado: [[0, 4, 12], [0, 0, 2], [7, 4, 6]]
   ACCIO: 17
   estado: [[0, 4, 12], [0, 3, 2], [7, 4, 6]]
   ACCIO: 17
   estado: [[0, 4, 12], [0, 6, 2], [7, 4, 6]]
   ACCIO: 2
   estado: [[0, 4, 12], [0, 6, 2], [7, 5, 6]]
   ACCIO: 1
   estado: [[0, 4, 12], [0, 6, 2], [6, 6, 6]]
   ACCIO: 7

estado: [[0, 4, 12], [0, 6, 2], [5, 5, 6]]
ACCIO: 0
estado: [[0, 4, 12], [0, 6, 2], [4, 5, 6]]
ACCIO: 0
estado: [[0, 4, 12], [0, 6, 2], [3, 5, 6]]
ACCIO: 1
estado: [[0, 4, 12], [0, 6, 2], [2, 6, 6]]
ACCIO: 7
estado: [[0, 4, 12], [0, 6, 2], [1, 5, 6]]
ACCIO: 5

Com podem observar els moviments de les peces del mariner son més erràtics, ja que no sempre fa l'acció que tria per culpa del seu estat d'embriaguesa. No sempre seguirà el camí més òptim.

d.) Compare the application of Q-learning in this chess scenario with that of the grid of exercise 1. How do the two scenarios differ? How does that translate into your results?

A la graella de l'exercici 1 hi havia molts menys estats possibles. La dificultat més grossa ha estat pensar com representar els possibles estats i moviments de les dues peces a moure. En conseqüència, la q-table del *chess* és molt més gran. La manera que hem usat per definir els estats ha estat 64 * 64 possibles posicions de les dues peces (files) per 36 accions (columnes). El rei pot fer 8 accions a cada moviment, depenent de la seva situació, i la torre en pot fer 24.
Al haver-hi tantes caselles a la q-table, el criteri de convergència és més sensible, i per tant trobem la resposta òptima en menys temps.

e.) Compare the use of Q-learning with that of search algorithms of P1 for the chess scenario seen here, both in the deterministic and stochastic case.

En els escenaris deterministes sí que seria possible aplicar els algorismes de la P1, ja que són algorismes enfocats en el determinisme, com l'A*. En el moment que l'estocàstica juga un paper més important no podem determinar el comportament de l'agent i no seria adient plantejar-ho amb els algorismes de la P1.

ANNEX FOTOS Q-TABLE:

```
PRIMERA Q-TABLE

[-inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, -0.4, -0.4, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, -inf, -inf, -0.4, -0.4, -0.4, -inf, -inf, -inf, -inf

Q-TABLE INTERMITJA

[-inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, -0.4, 0, 0, 0, -0.4, -inf, -inf, -inf, -inf, -inf, -
[-inf, -inf, 10.50361962368071, 0, 0, 0, -0.64, -inf, -inf, -inf,
[-inf, -inf, -0.4, 47.38921721393924, 0, -0.4, 1.7007239247361419
[-inf, -inf, -0.4, -0.4, -0.4, -0.4, 27.059385049189142, -inf, -i
[-inf, -inf, -inf, -inf, -0.4, -0.64, -0.4, -inf, -inf, -inf, -in
```

```
SEGONA Q-TABLE INTERMITJA

[-inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -in
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf,
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf,
[-inf, -inf, -0.4, 0, 0, 0, -0.4, -inf, -inf, -inf, -inf, -inf, -inf, -inf,
[-inf, -inf, 10.50361962368071, 0, 0, 0, -0.64, -inf, -inf, -inf, -inf, -inf
[-inf, -inf, -0.4, 47.38921721393924, 0, -0.4, 1.7007239247361419, -inf, -in
[-inf, -inf, -0.4, -0.4, -0.4, -0.4, 27.059385049189142, -inf, -inf, -inf, -
[-inf, -inf, -inf, -inf, -0.4, -0.64, -0.4, -inf, -inf, -inf, -inf, -inf, -i

°Q-TABLE FINAL

[-inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -in
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf,
[-inf, -inf, 0, 0, 0, 0, 0, -inf, -inf, -inf, -inf, -inf, -inf, -inf, -inf,
[-inf, -inf, -0.4, 0, 0, 0, -0.4, -inf, -inf, -inf, -inf, -inf, -inf, -inf,
[-inf, -inf, 10.50361962368071, 0, 0, 0, -0.64, -inf, -inf, -inf, -inf, -inf
[-inf, -inf, -0.4, 47.38921721393924, 0, -0.4, 1.7007239247361419, -inf, -in
[-inf, -inf, -0.4, -0.4, -0.4, -0.4, 27.059385049189142, -inf, -inf, -inf, -
[-inf, -inf, -inf, -inf, -0.4, -0.64, -0.4, -inf, -inf, -inf, -inf, -inf, -i
```