سیستم های هوشمند

دکتر رشاد حسینی

گزارش تمرین کامپیوتری ۴

پوریا آزادی مقدم

۸۱۰۱۹۳۳۳۱

نحوه ساختن ماتریس R :

این ماتریس نشانده امتیاز گرفته شده در اثر رفتن از یک state به وسیله عملی خاص به state دیگر میباشد به همین منظور در ابتدا لازم است ساختار state توضیح داده شود:

- هر state مشخص کننده وضعیت هر یک از دیسک ها بر روی میله ها میباشد, توجه شود که 3 میله داریم که با شماره 0 ,1 , 2 نامگذاری شده اند.

- حال هر state یک وکتور از اعداد صحیح 0 تا 2 میباشد که طول آن برابر با تعداد دیسک ها یعنی n میباشد.

- در این وکتور دیسک های بزرگتر دارای ایندکس کوچتر و دیسک های کوچمتر دارای ایندکس بزرگتر میباشد, یعنی وضعیت بزرگترین دیسک را ایندکس 1 این وکتور و وضعیت قرارگیری کوچکترین دیسک یعنی دیسک n ام برابر با ایندکس n میباشد.

در شکل صفحه بعد میتوانید state های موجود برای 3 حالت 3 دیسک را مشاهده نمایید.

توجه شود که ایندکس n +1 ام برابر با شماره هر استیت در ماتریس states که شامل تمام استیت ها است, میباشد.

	1	2	3	4
1	0	0	0	1
2	0	0	1	2
3	0	0	2	3
4	0	1	0	4
5	0	1	1	5
6	0	1	2	6
7	0	2	0	7
8	0	2	1	8
9	0	2	2	9
10	1	0	0	10
11	1	0	1	11
12	1	0	2	12
13	1	1	0	13
14	1	1	1	14
15	1	1	2	15
16	1	2	0	16
17	1	2	1	17
18	1	2	2	18
19	2	0	0	19
20	2	0	1	20
21	2	0	2	21
22	2	1	0	22
23	2	1	1	23
24	2	1	2	24
25	2	2	0	25
26	2	2	1	26
27	2	2	2	27

شکل 1- استیت ها به همراه شماره ایندکس

همان طور که مشخص است 2^n -1 **استیت** داریم که حالت ممکنی برای قرار گیری این n دیسک را نشان میدهد.

توضیح ماتریس R :

در این ماتریس سطر ها نشانده state حال حاضر است و ستون ها نیز نشان دهنده استیت های مرحله بعد است که در اثر یک عمل خاص به ان خواهیم رفت.

در ساخت این ماتریس باید به نکات زیر که قوانین بازی است توجه نمود:

- در هر زمان فقط یک دیسک را میتوان جابهجا نمود.
- نباید در هیچ زمانی دیسکی بر روی دیسکی کوچکتر قرار گیرد.

دو قانون بالا نشان دهنده ان است که رفتن از هر استیت به استیت های دیگر 3 حالت دارد:

- رفتن از استیت حال حاضر به استیت بعدی مورد نظر بر اساس قوانین بالا مجاز باشد اما استیت بعدی, مرحله نهایی نباشد که امتیاز این حرکت برابر با 0.01- خواهد بود.

- رفتن از استیت حال حاضر به استیت بعدی مورد نظر بر اساس قوانین بالا ناممکن و مجاز نباشد , امتیاز این حالت برابر با منفی بینهایت یا عبارتی منفی بزرگی خواهد بود.

- رفتن از استیت حال حاضر به استیت بعدی مورد نظر بر اساس قوانین بالا مجاز باشد اما استیت بعدی, مرحله نهایی نباشد که امتیاز این حرکت برابر با 100 خواهد بود.

ماتریس های Qو R در فایل ارسالی موجود است.

شکل زیر نشاندهده ماتریس R برای حالت 3 دیسک خواهد بود:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
2	-0.0100	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
3	-0.0100	-0.0100	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
4	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
5	-Inf	-Inf	-Inf	-0.0100	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
6	-Inf	-Inf	-0.0100	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
7	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf
8	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf
9	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
10	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-Inf	-Inf
11	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-0.0100	-Inf	-Inf
12	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf
13	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100
14	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf
15	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-0.0100
16	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf
17	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf
18	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf
19	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf
20	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
21	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
22	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
23	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
24	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
25	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
26	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
27	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf

شکل2- بخش اول ماتریس R

15	16	17	18	19	20	21	22	23	24	25	26	27
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-0.0100	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf
-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-0.0100	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-0.0100	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-0.0100	-Inf	-Inf	-Inf	-Inf
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-0.0100	100
-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	-Inf	-Inf	-Inf	-0.0100	-Inf	100
-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-Inf	-0.0100	-0.0100	-Inf

شکل2- بخش دوم ماتریسR

همچنین برای ماتریس Qخواهیم داشت:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	58.2100	72.7751	0	0	0	0	0	0	0	0	0	0	0
2	-0.0386	0	72.7751	0	0	0	0	-0.0371	0	0	0	0	0	0
3	58.2100	58.2100	0	0	0	90.9813	0	0	0	0	0	0	0	0
4	0	0	0	0	113.7392	-0.0310	-0.0345	0	0	0	0	0	0	0
5	0	0	0	90.9813	0	90.9722	0	0	0	0	0	0	0	0
6	0	0	72.7751	90.9813	113.7392	0	0	0	0	0	0	0	0	0
7	0	0	0	81.0527	0	0	0	-0.0358	-0.0345	0	0	0	0	0
8	0	-0.0351	0	0	0	0	-0.0347	0	-0.0350	0	0	0	0	0
9	0	0	0	0	0	0	-0.0355	-0.0323	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	-0.0263	-0.0278	0	0
11	0	0	0	0	0	0	0	0	0	-0.0277	0	-0.0274	0	0
12	0	0	0	0	0	0	0	0	0	-0.0274	-0.0286	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.0264
14	0	0	0	0	0	0	0	0	0	0	0	0	-0.0273	0
15	0	0	0	0	0	0	0	0	0	0	0	-0.0283	-0.0278	-0.0263
16	0	0	0	0	0	0	0	0	0	0	0	0	-0.0278	0
17	0	0	0	0	0	0	0	0	0	0	-0.0283	0	0	0
18	0	0	0	0	0	0	0	0	-0.0314	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	-0.0218	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	113.7392	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0

شکل2- بخش اول ماتریسQ

14	15	16	17	18	19	20	21	22	23	24	25	26	27
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	142.1864	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	-0.0327	0	0	0	0	0	0	0	0	0
0	0	0	0	0	-0.0231	0	0	0	0	0	0	0	0
0	0	0	-0.0277	0	0	0	0	0	0	0	0	0	0
0	-0.0302	0	0	0	0	0	0	0	0	0	0	0	0
-0.0264	-0.0275	-0.0236	0	0	0	0	0	0	0	0	0	0	0
0	-0.0240	0	0	0	0	0	0	0	0	0	0	0	0
-0.0263	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	-0.0284	-0.0347	0	0	0	0	0	0	0	0	0
0	0	-0.0283	0	-0.0306	0	0	0	0	0	0	0	0	0
0	0	-0.0319	-0.0322	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	46.6382	-0.0224	0	0	0	0	0	0
0	0	0	0	0	-0.0164	0	-0.0164	0	0	0	0	220.6204	0
0	0	0	0	0	-0.0178	-0.0178	0	0	0	102.3636	0	0	0
0	0	0	0	0	0	0	0	0	142.1864	113.7392	222.1944	0	0
0	0	0	0	0	0	0	0	177.7456	0	113.7392	0	0	0
0	0	0	0	0	0	0	-0.0230	-0.0164	142.1864	0	0	0	0
0	0	0	0	0	0	0	0	177.7456	0	0	0	222.1944	277.7556
0	0	0	0	0	0	148.6360	0	0	0	0	199.9741	0	277.7556
0	0	0	0	0	0	0	0	0	0	0	222.1944	222.1944	0

شکل2- بخش دوم ماتریسQ

سوال 2: نحوه تبدیل ماتریس R به ماتریس Q

توجه شود که Q نیز ماتریسی به ابعاد $3^{n\ discs}$ می باشد .

در ابتدا این ماتریس با 0 مقدار دهی اولیه میشود ,سپس با exploration و exploitation توسط ایجنت بین state های مختلف, این ماتریس اپدیت خواهد شد:

- برای اپدیت کردن و ساختن ماتریس Q به این صورت عمل خواهیم کرد که در هر اپیزود از یک استیت مشخص شروع خواهیم کرد, این استیت مشخص اولیه میتواند به صورت رندم انتخاب شود یا برای ان که ذات بازی کردن ایجنت حفظ شود, از مرحله استیت 1 یعنی جایی که تمام دیسک ها بر روز میله 0 قرار دارند شروع شود.

- با هر حرکت مجاز ایجنت, آن را از استیت حال به استیت دیگری میبرد, ایجنت در اثر این حرکت مقداری را به عنوان reward دریافت خواهد کرد که مقدار این reward از طریق ماتریس Rبا نگاه کردن به درایه موجود در سطر current state و ستون next state بدست خواهد امد.

- همان طور که ذکر شد این مقدار reward برای حرکات مجاز که منجر به بردن بازی نمیشود برابر با 0.01- و اگر منجر به برد شود برابر با 100 خواهد بود.

- توجه نمایید که برای خاصیت exploration و exploitation مربوط به ایجنت, نیاز به یک policy داریم. در این شبیه سازی از سیاست epsilon greedy استفاده شده است.

- در این سیاست مقداری به عنوان اپسیلون بین 0 تا 1 انتخاب میشود, با احتمال 1-epsilon ایجنت به استیت ممکن بعدی با بیشترین مقدار Qخواهد رفت و با احتمال epsilon به صورت رندم از استیت های ممکن جهت حرکت بعدی خود را مشخص خواهد کرد.

- هرچه ایجنت بیشتر در بین استیت ها حرکت نماید مقدار Q بهتر ای بدست خواهد امد.

- در اپدیت کردن مقدار Qعلاوه بر در نظر گرفتن مقدار reward که نگاه به حرکت کنونی دارد , به اینده نیز میتوان نگاه کرد و بر اساس پیش بینی انجام شده مقدار متناسبی را برای اپدیت کردن Q در نظر گرفت.

- برای این اپدیت کردن به استیت انتخاب شده به وسیله epsilon greedy به عنوان next state ,نگاه خواهیم کرد. بیشترین Q ممکن که ایجنت میتواند در استیت بعد از next state , بدست بیاورد را میتوان در ضریبی ,ضرب کرده تا اثر اینده و پیشبینی را وارد نمود.

با توجه این موارد ذکر شده , میتوان از رابطه زیر جهت اپدیت کردن ماتریس Q استفاده نمود:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(R(s) + \beta \max_{a'} Q(s',a') - Q(s,a)\right)$$

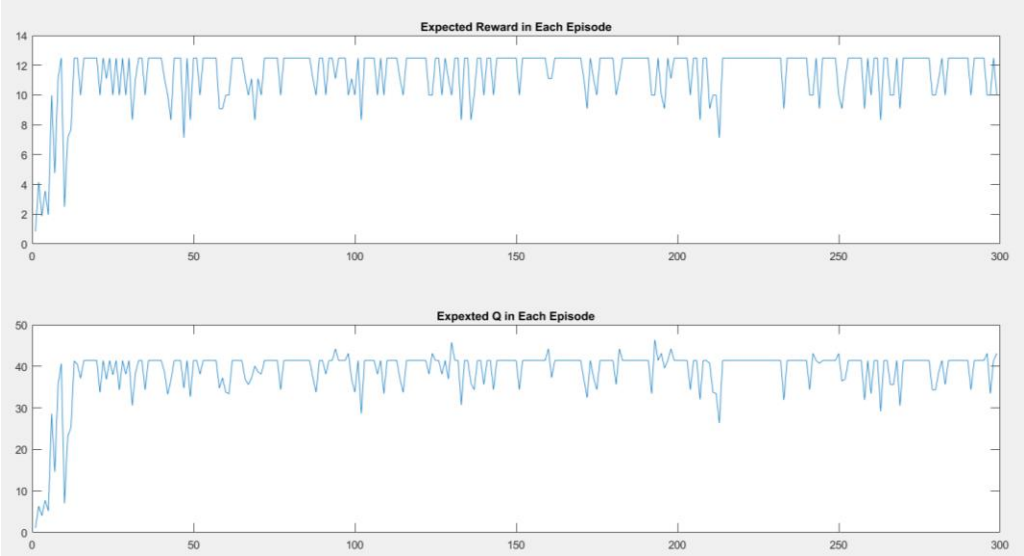
در فرمول بالا B.max_Q(s’,a’) همان نگاه ب اینده است که B را discount factor در نظر گرفته اند.

همچنین الفا learning rate میباشد.

این دو پارمتر در سوال 3 شرح داده خواهند شد.

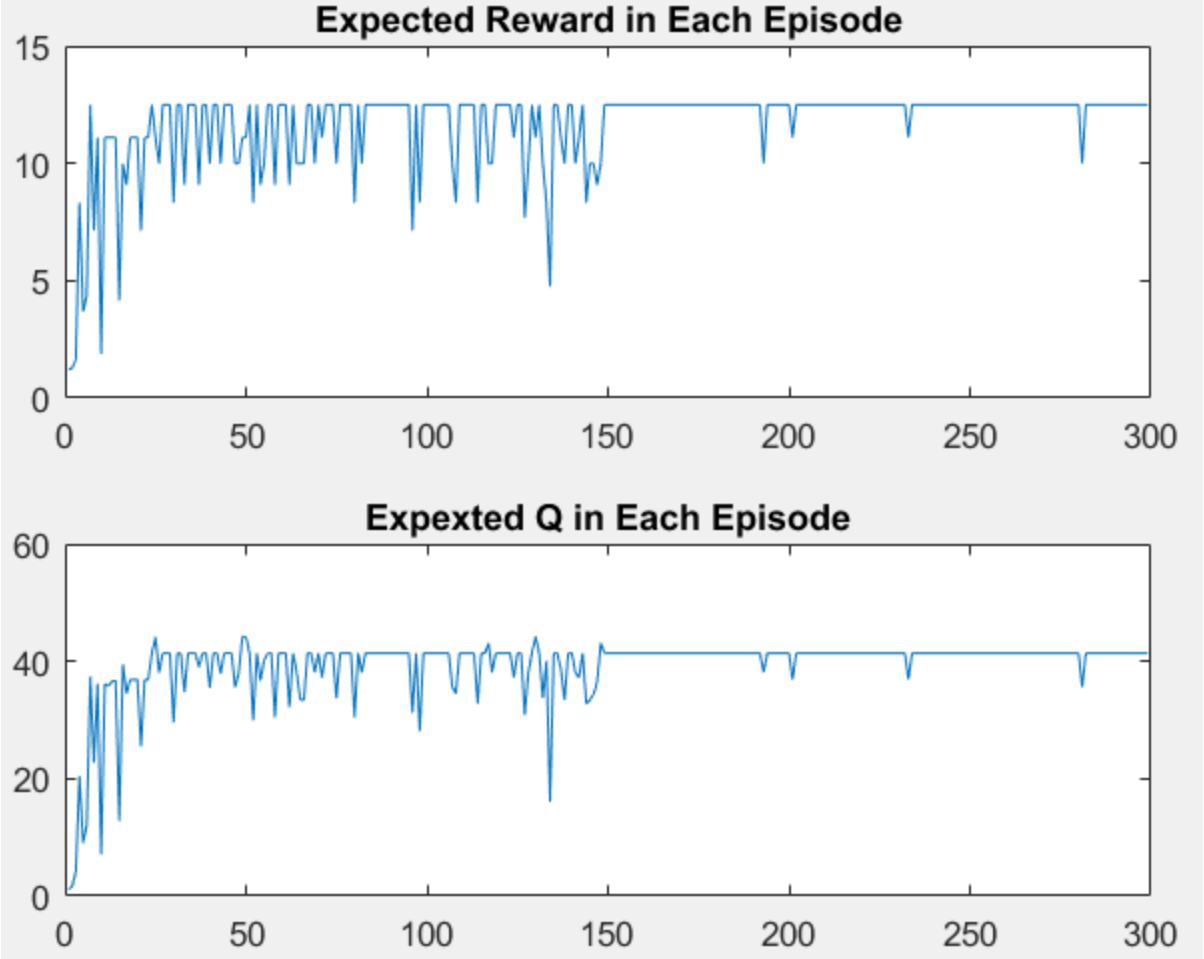
سوال 3:

برای 3 دیسک مقدار متوسط reward و Qجمع اوری شده به صورت زیر میشود:



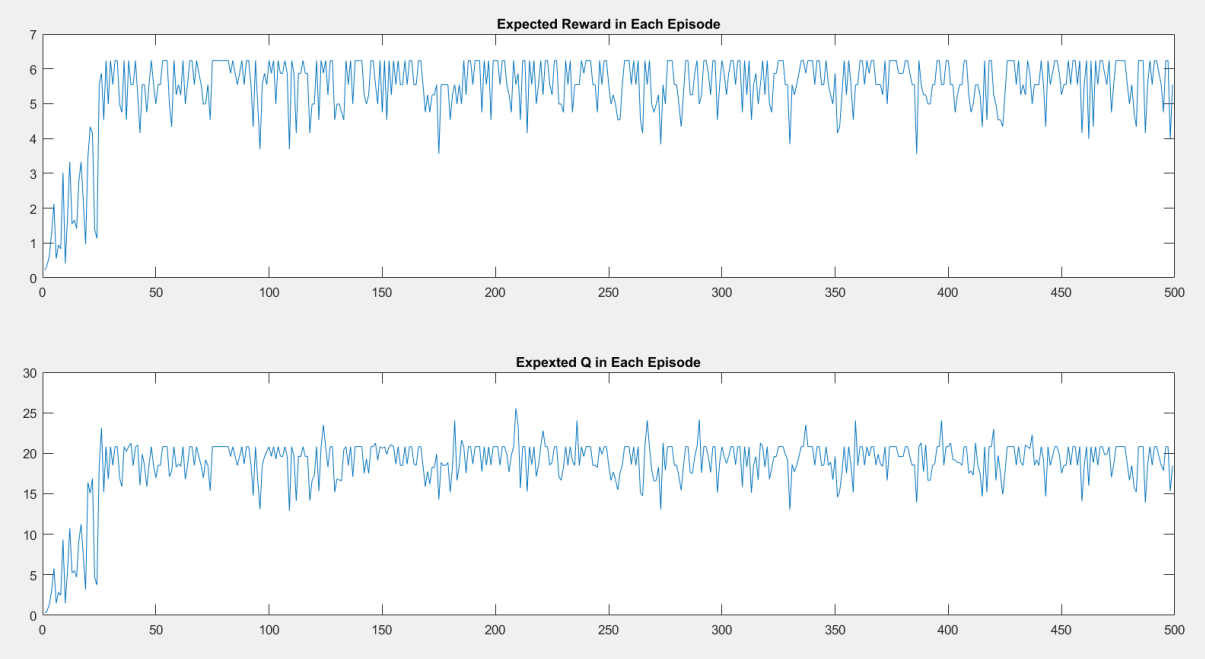
شکل 3- مقدار متوسط امتیاز جمع اوری شده برای 3 دیسک

همچنین اگر مقدار اپسیلون را از اپیزود 150 به بعد به مقدار0.01 کاهش دهیم نمودار ها به شکل زیر میشود:



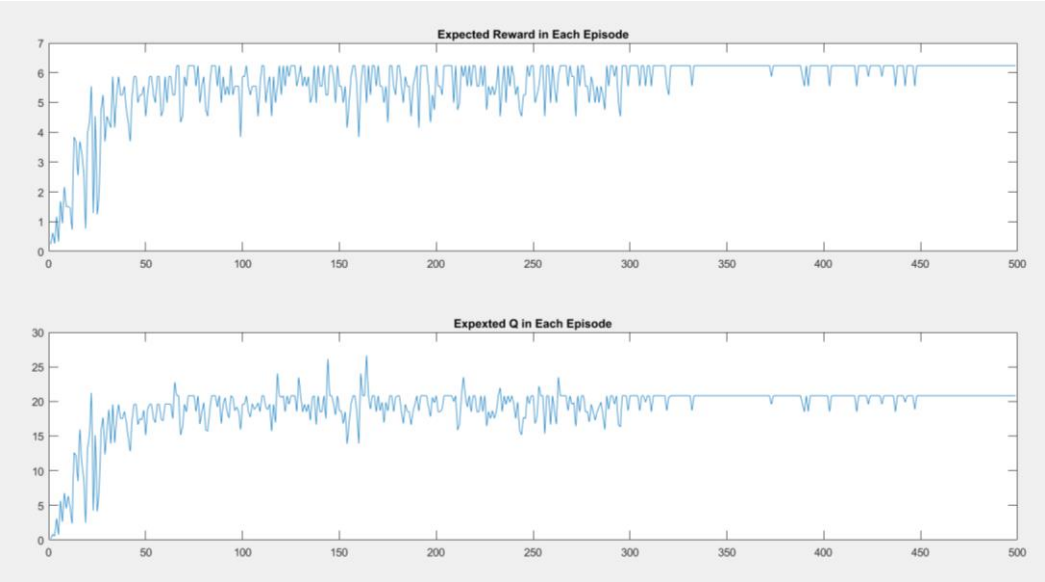
شکل 4- مقدار متوسط امتیاز جمع اوری شده برای 3 دیسک

برای 4 دیسک مقدار متوسط reward و Qجمع اوری شده به صورت زیر میشود:



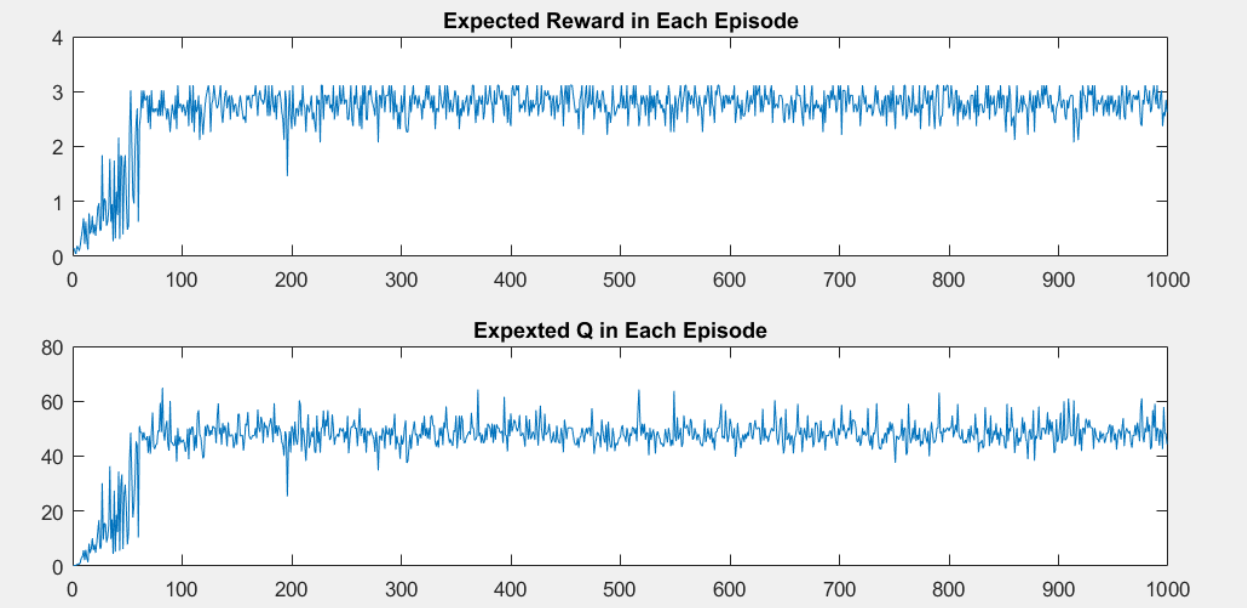
شکل 5- مقدار متوسط امتیاز جمع اوری شده برای 4 دیسک

همچنین اگر مقدار اپسیلون را از اپیزود 300 به بعد به مقدار0.01 کاهش دهیم نمودار ها به شکل زیر میشود:



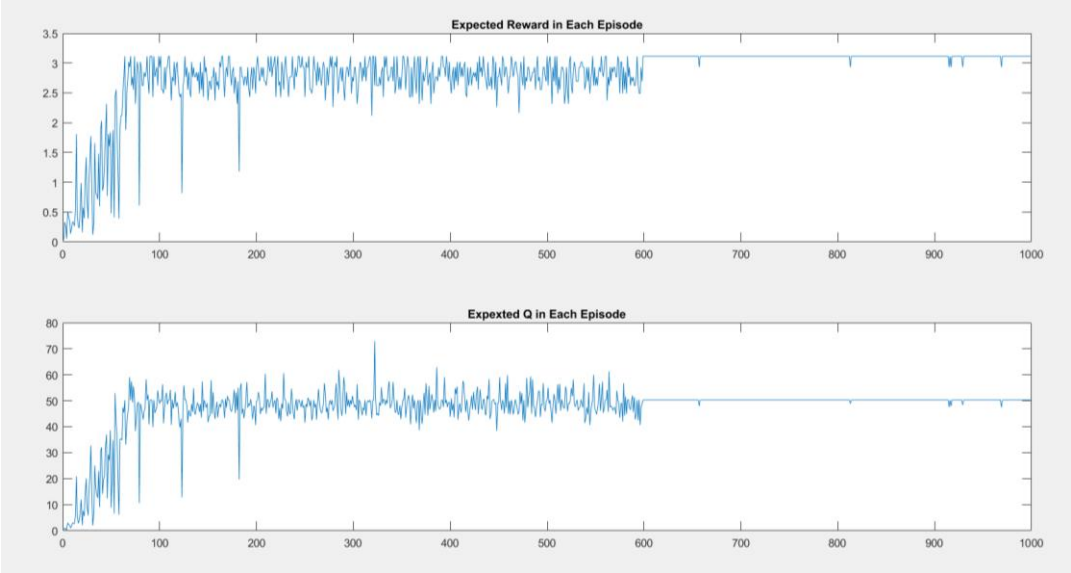
شکل 6- مقدار متوسط امتیاز جمع اوری شده برای 4 دیسک

برای 5 دیسک مقدار متوسط reward و Qجمع اوری شده به صورت زیر میشود:



شکل 7- مقدار متوسط امتیاز جمع اوری شده برای 4 دیسک

همچنین اگر مقدار اپسیلون را از اپیزود 600 به بعد به مقدار0.001 کاهش دهیم نمودار ها به شکل زیر میشود:

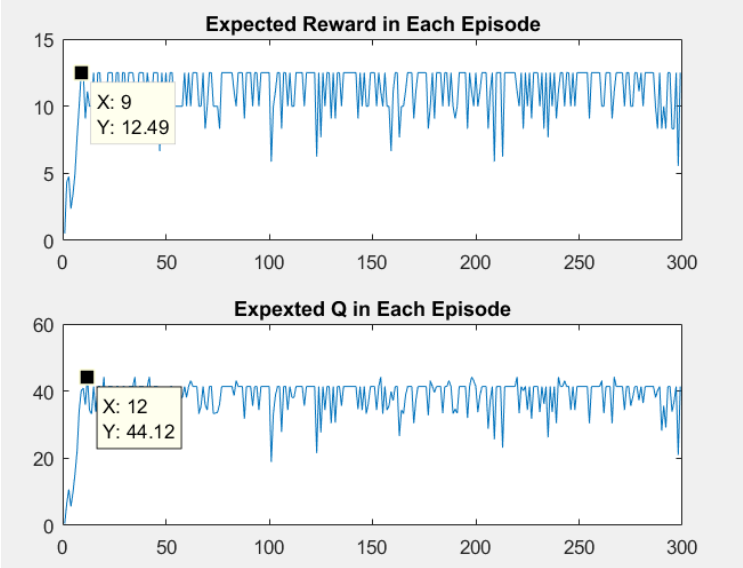


شکل 8- مقدار متوسط امتیاز جمع اوری شده برای 5 دیسک

بررسی اثر نرخ یادگیری:

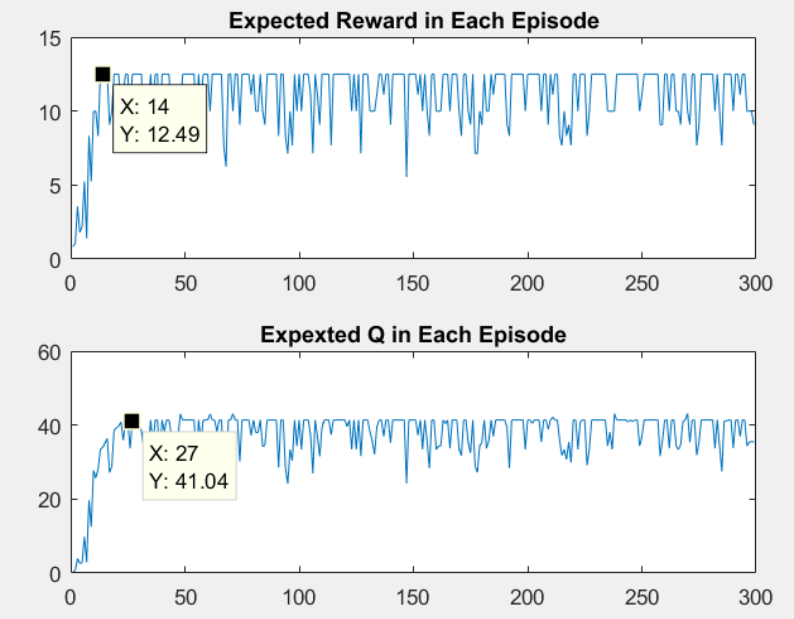
برای 3 دیسک داریم:

در ابتدا نرخ یادگیری را برابر 1 قرار خواهیم داد:



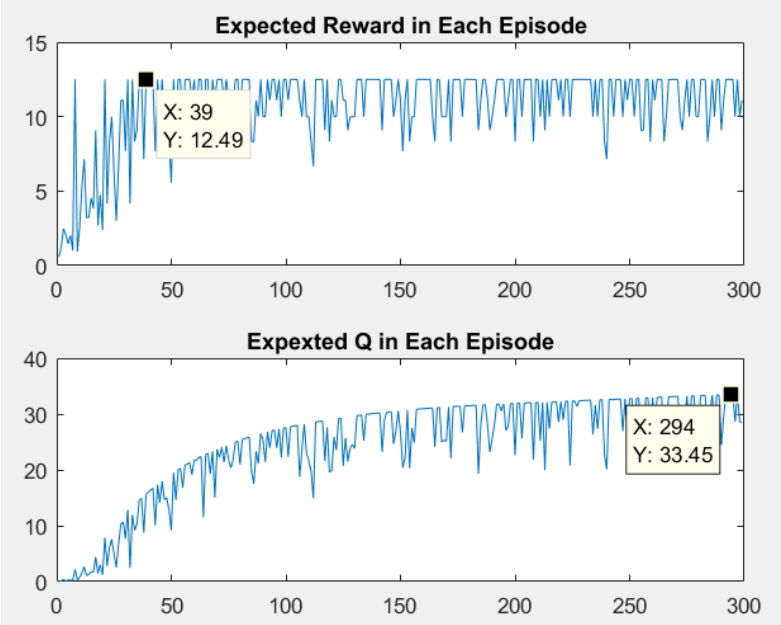
شکل9-میانگین جایزه ها با 3 دیسک و الفا برابر 1

برای نرخ 0.5 داریم:



شکل 10-میانگین جایزه ها با 3 دیسک و الفا برابر 0.05

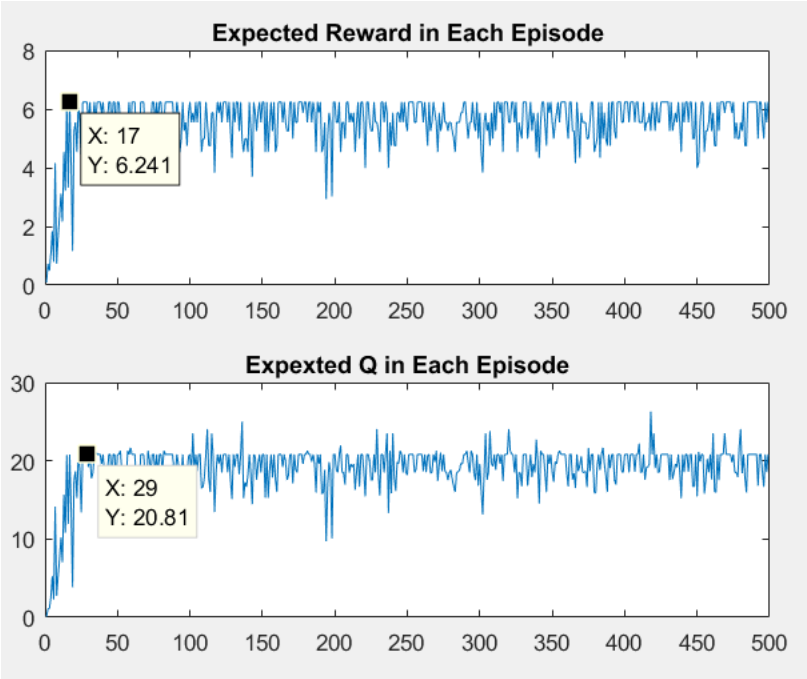
نرخ یادگیری 0.05:



شکل 11-میانگین جایزه ها با 3 دیسک و الفا برابر 0.05

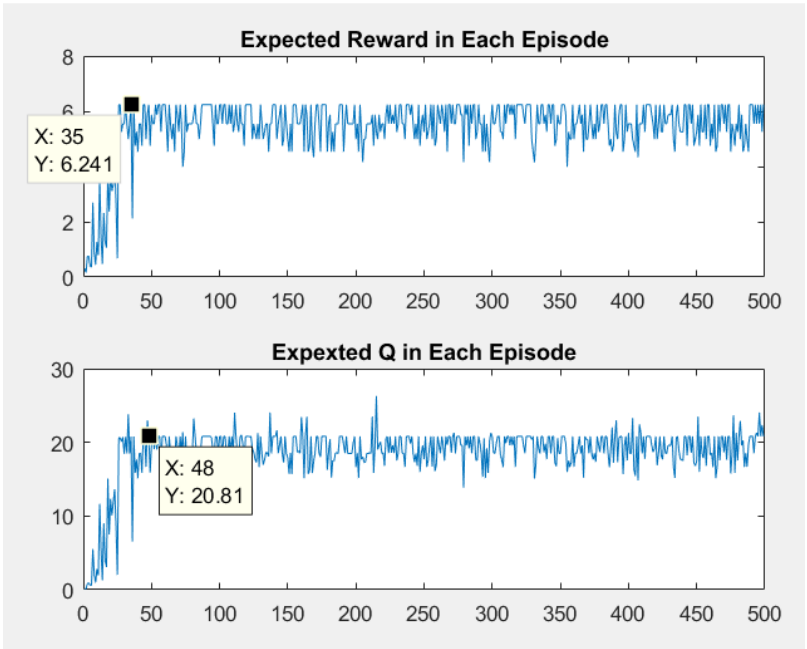
برای 4 دیسک داریم:

در ابتدا نرخ یادگیری را برابر 1 قرار خواهیم داد:



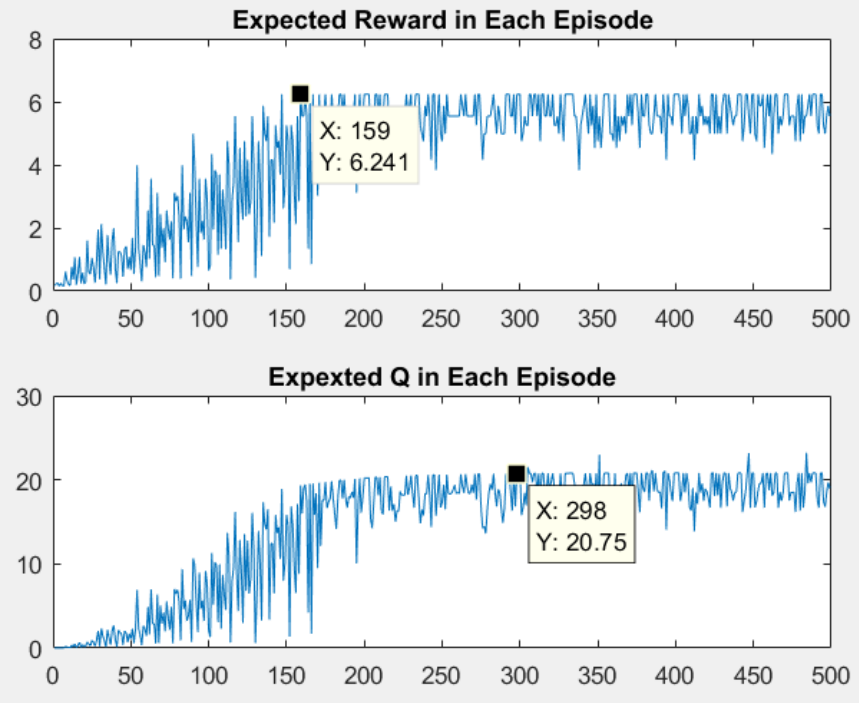
شکل12-میانگین جایزه ها با 4 دیسک و الفا برابر 1

برای نرخ 0.5 داریم:



شکل 13-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

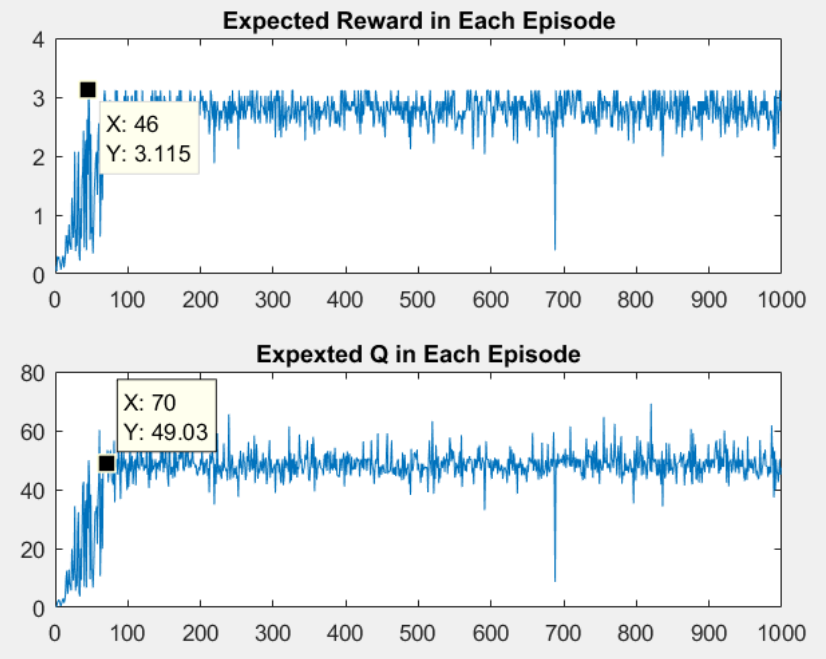
نرخ یادگیری 0.05:



شکل 14-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

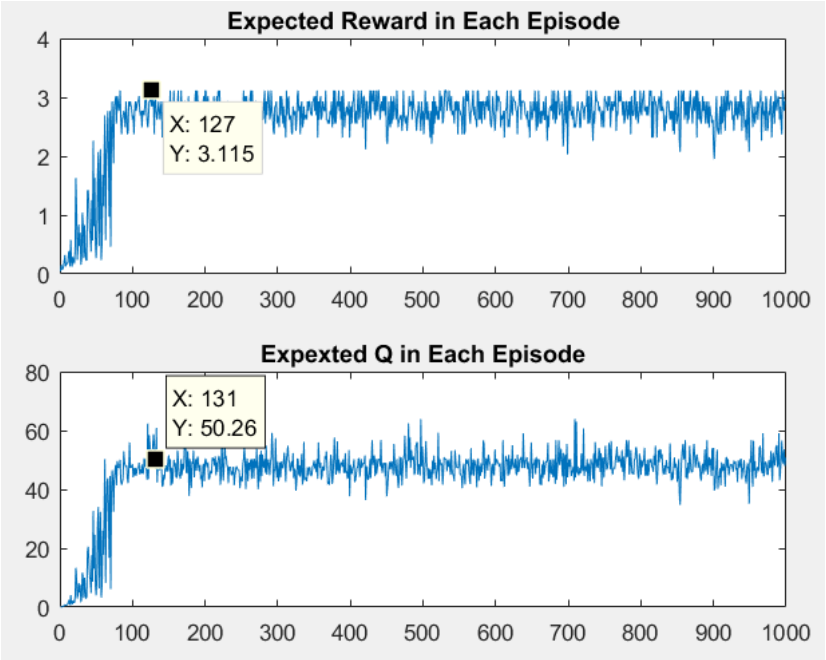
برای 5 دیسک داریم:

در ابتدا نرخ یادگیری را برابر 1 قرار خواهیم داد:



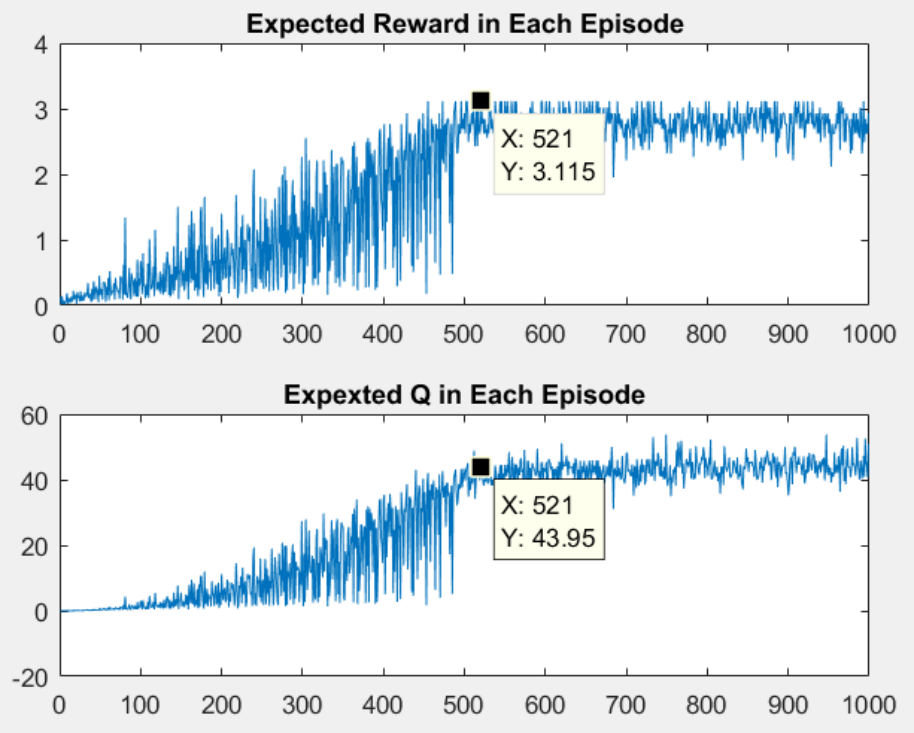
شکل15-میانگین جایزه ها با 4 دیسک و الفا برابر 1

برای نرخ 0.5 داریم:



شکل 16-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

نرخ یادگیری 0.05:



شکل 17-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

نتایج مقایسه شکل های بالا بر حسب تعداد episode ها در جدول 1 میتوانید مشاهده نمایید:

الفا/تعداد دیسک	3 دیسک	4 دیسک	5دیسک
الفا =1	12	29	70
الفا=0.5	27	48	131
الفا=0.05	294	298	528

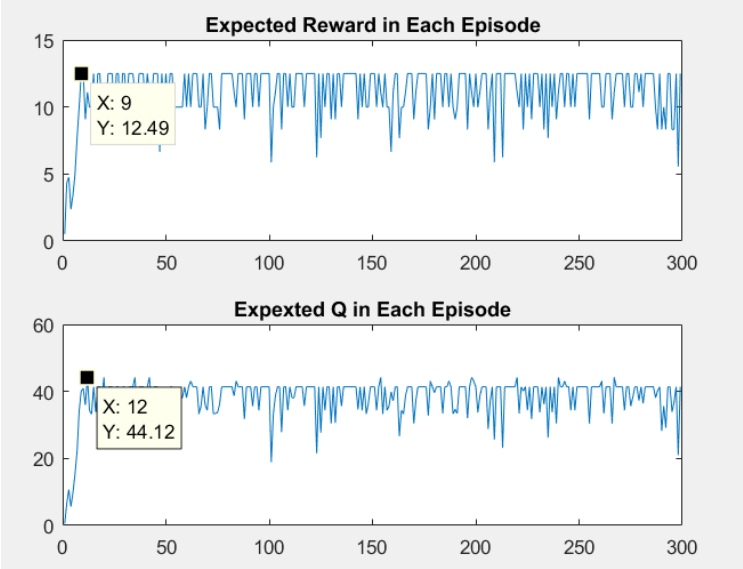
جدول 1- مقایسه تعدادمراحل برای دیسک های مختلف به ازای الفا های متفاوت

همان طور که مشاهده می نمایید الفا یا همان نرخ یادگیری سرعت رشد و یادگیری agent را مشخص میکند. هر چه این پارامتر بزرگتر باشد اثر جایزه همان لحظه بیشتر شده و وابستگی به مقدار قبلی Q کمتر خواهد شد پس در بخش های اولیه یادگیری که مقادیر نزدیک هم میباشند, سرعت بسیار قابل تغییر بر اساس این پارامتر خواهد بود.

بررسی اثر نرخ یادگیری:

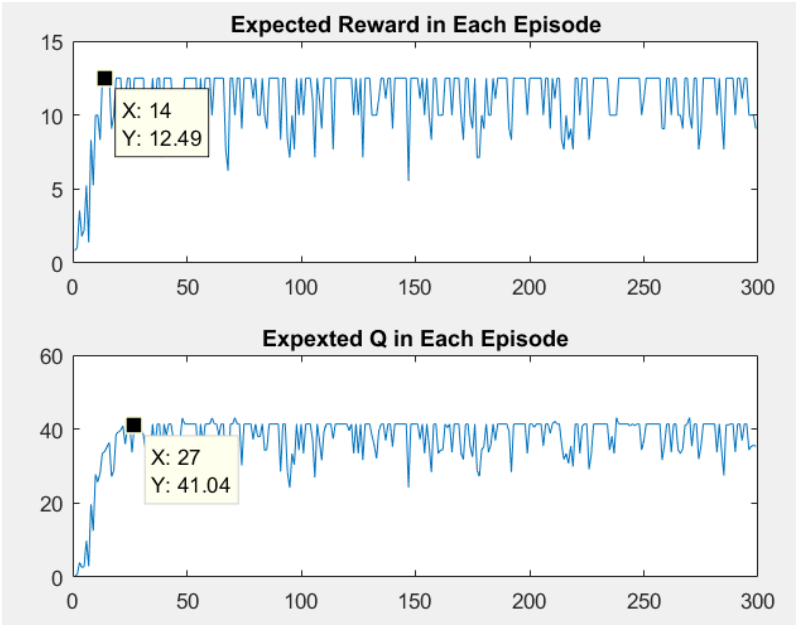
برای 3 دیسک داریم:

در ابتدا نرخ یادگیری را برابر 1 قرار خواهیم داد:



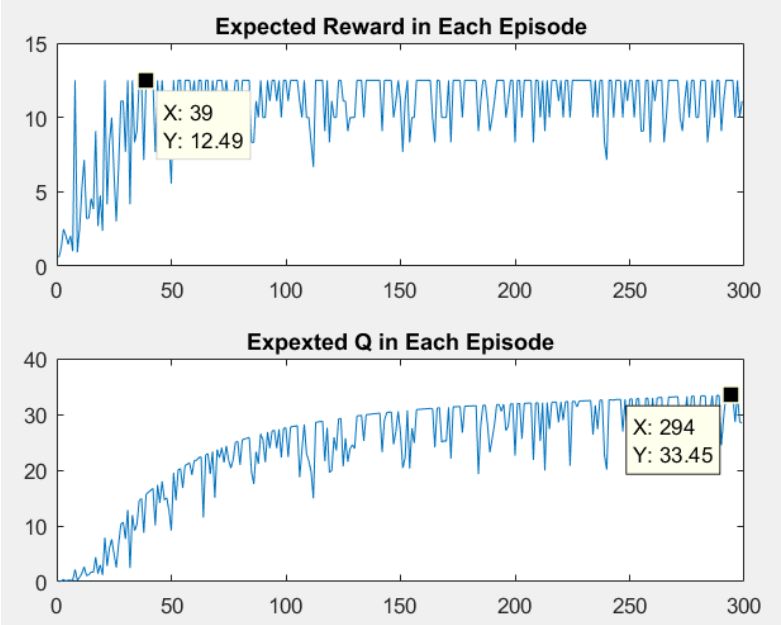
شکل9-میانگین جایزه ها با 3 دیسک و الفا برابر 1

برای نرخ 0.5 داریم:



شکل 10-میانگین جایزه ها با 3 دیسک و الفا برابر 0.05

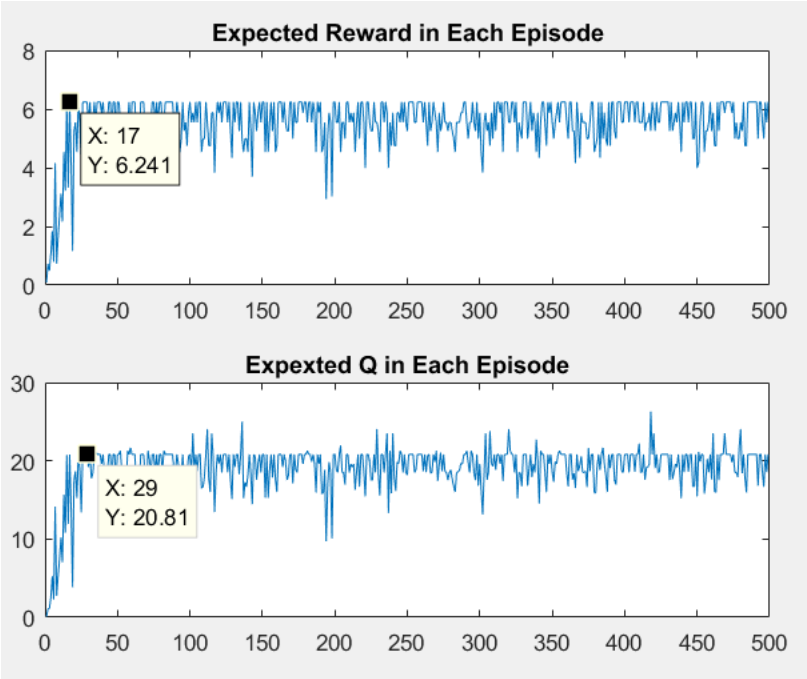
نرخ یادگیری 0.05:



شکل 11-میانگین جایزه ها با 3 دیسک و الفا برابر 0.05

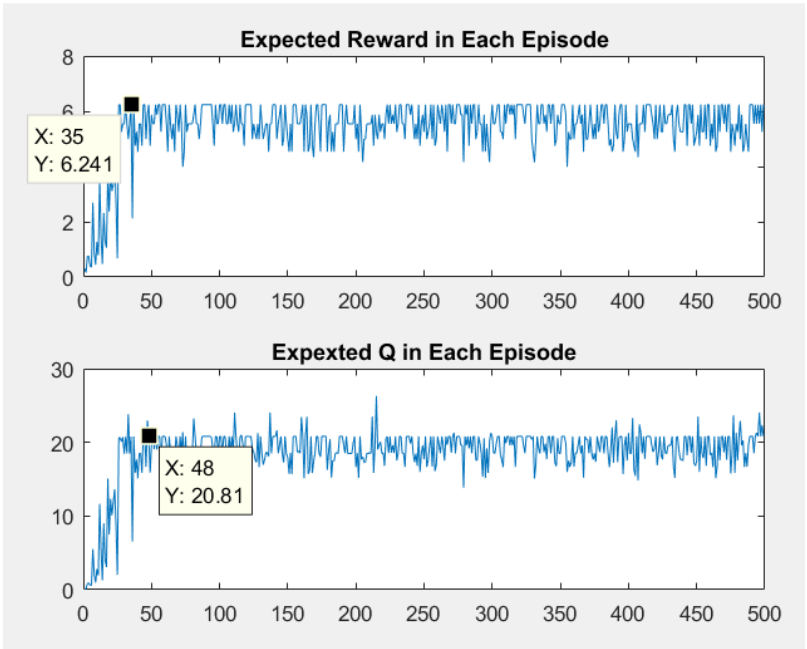
برای 4 دیسک داریم:

در ابتدا نرخ یادگیری را برابر 1 قرار خواهیم داد:



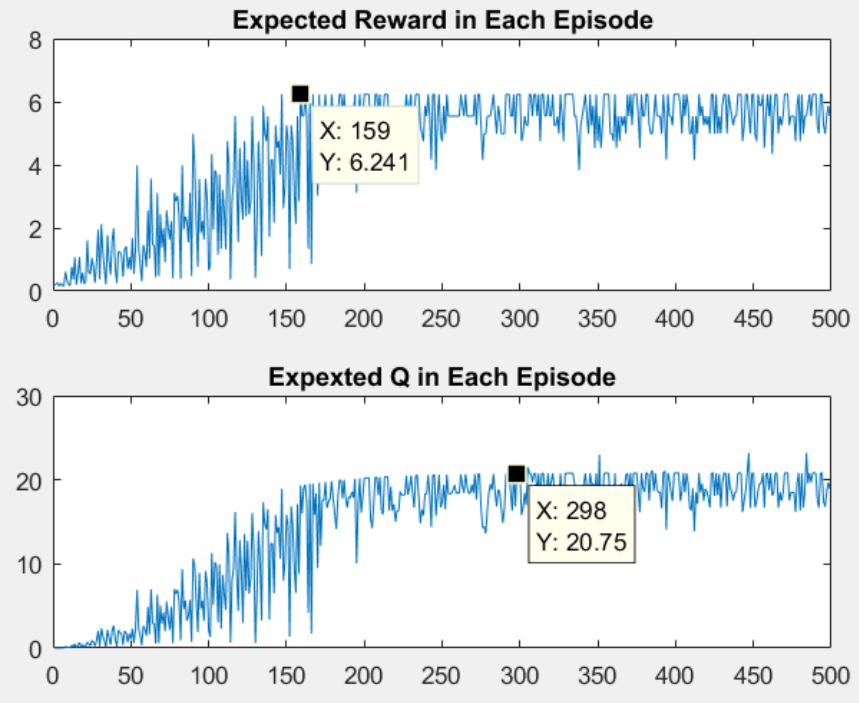
شکل12-میانگین جایزه ها با 4 دیسک و الفا برابر 1

برای نرخ 0.5 داریم:



شکل 13-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

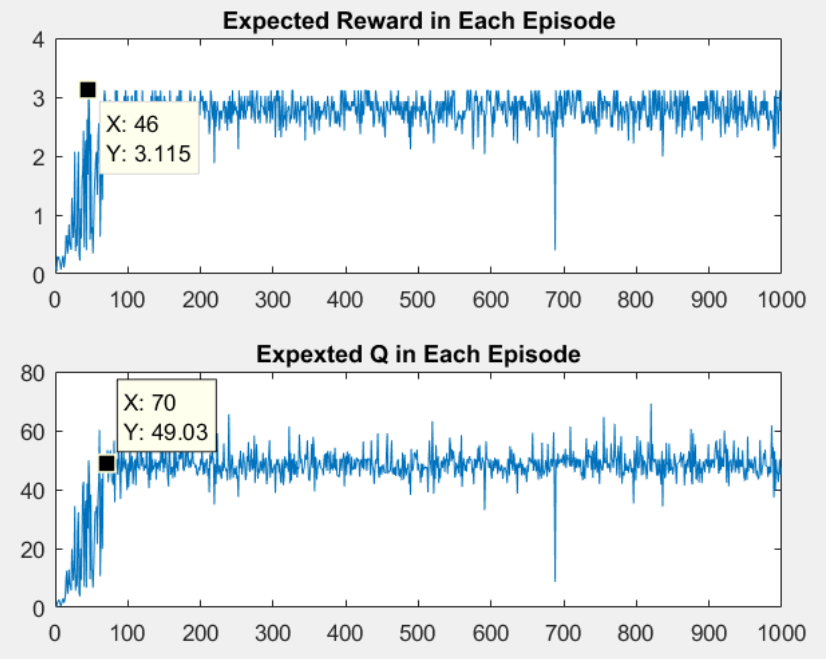
نرخ یادگیری 0.05:



شکل 14-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

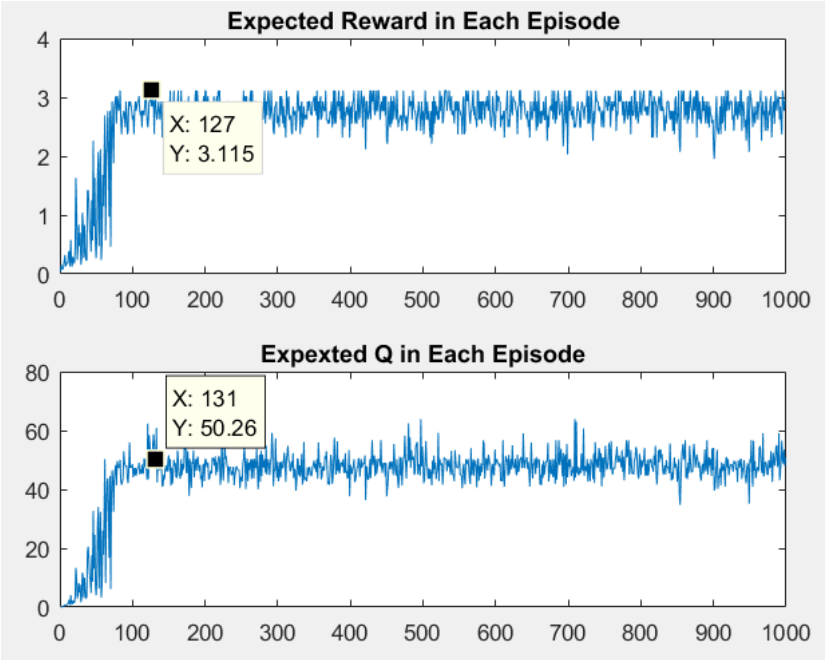
برای 5 دیسک داریم:

در ابتدا نرخ یادگیری را برابر 1 قرار خواهیم داد:



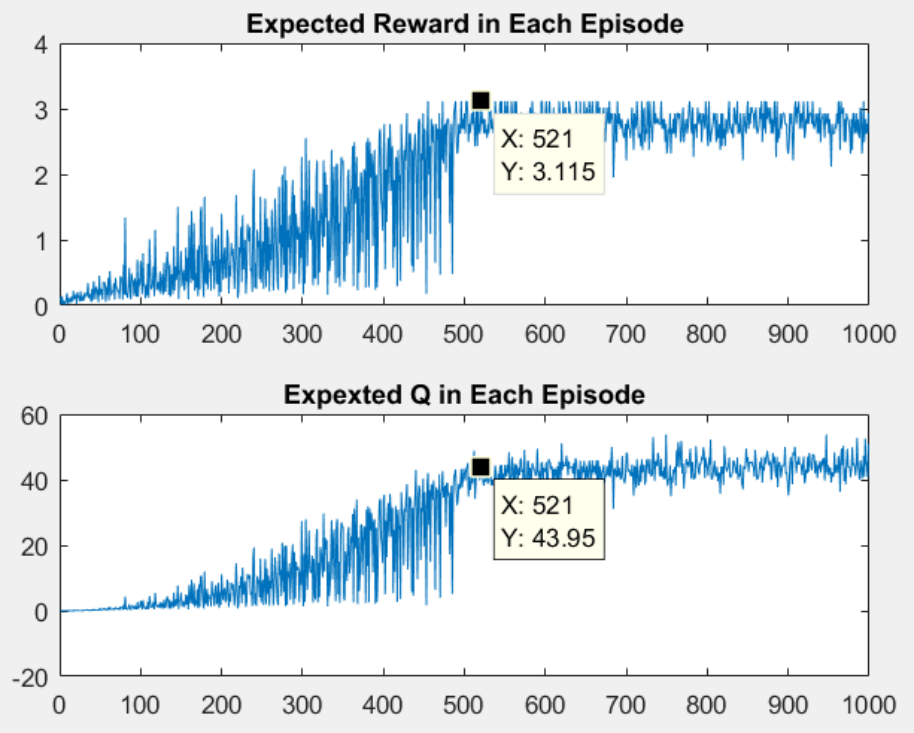
شکل15-میانگین جایزه ها با 4 دیسک و الفا برابر 1

برای نرخ 0.5 داریم:



شکل 16-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

نرخ یادگیری 0.05:



شکل 17-میانگین جایزه ها با 4 دیسک و الفا برابر 0.05

نتایج مقایسه شکل های بالا بر حسب تعداد episode ها در جدول 1 میتوانید مشاهده نمایید:

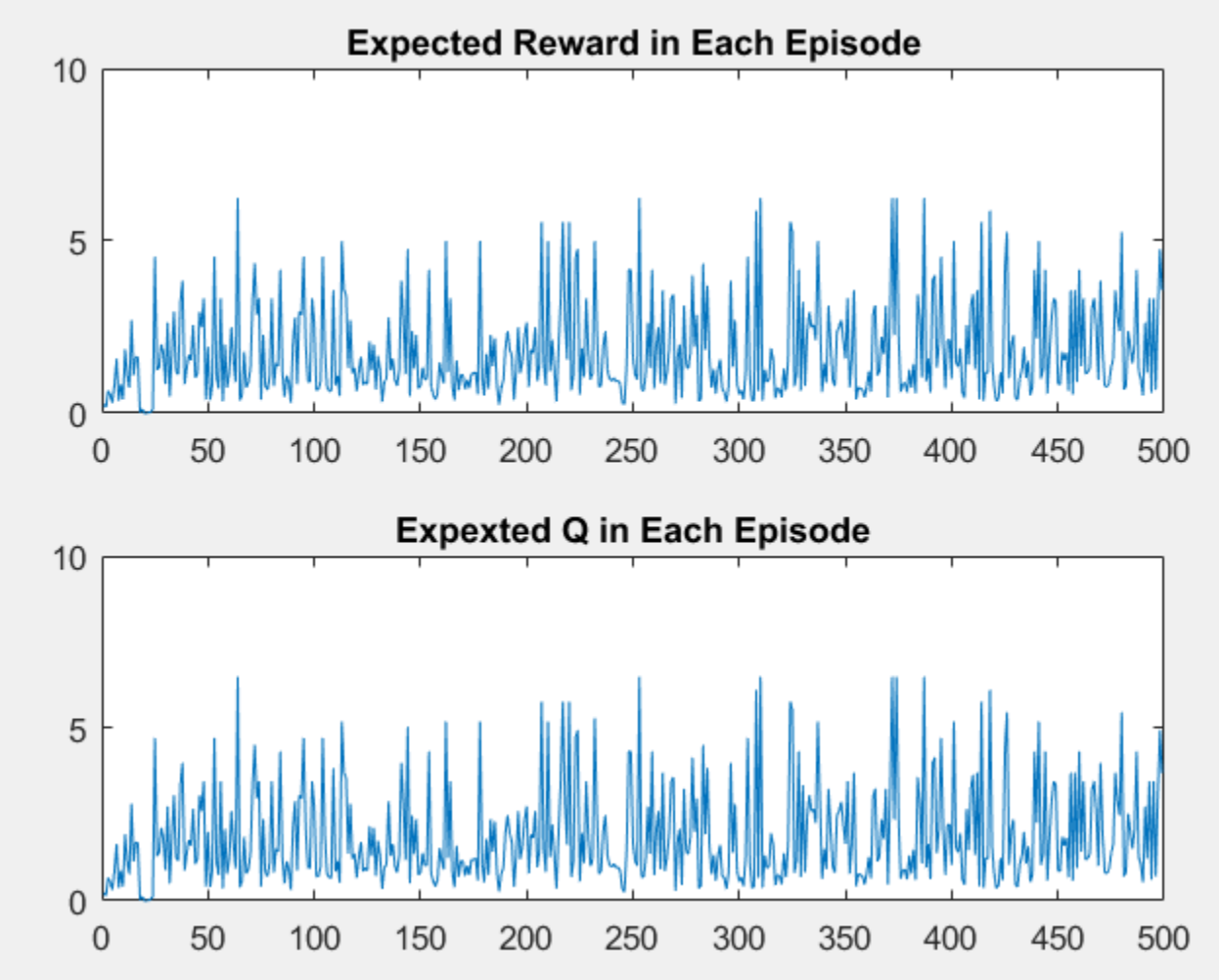
الفا/تعداد دیسک	3 دیسک	4 دیسک	5دیسک
الفا =1	12	29	70
الفا=0.5	27	48	131
الفا=0.05	294	298	528

جدول 1- مقایسه تعداد مراحل برای دیسک های مختلف به ازای الفا های متفاوت

همان طور که مشاهده می نمایید الفا یا همان نرخ یادگیری سرعت رشد و یادگیری agent را مشخص میکند. هر چه این پارامتر بزرگتر باشد اثر جایزه همان لحظه بیشتر شده و وابستگی به مقدار قبلی Q کمتر خواهد شد پس در بخش های اولیه یادگیری که مقادیر نزدیک هم میباشند, سرعت بسیار قابل تغییر بر اساس این پارامتر خواهد بود.

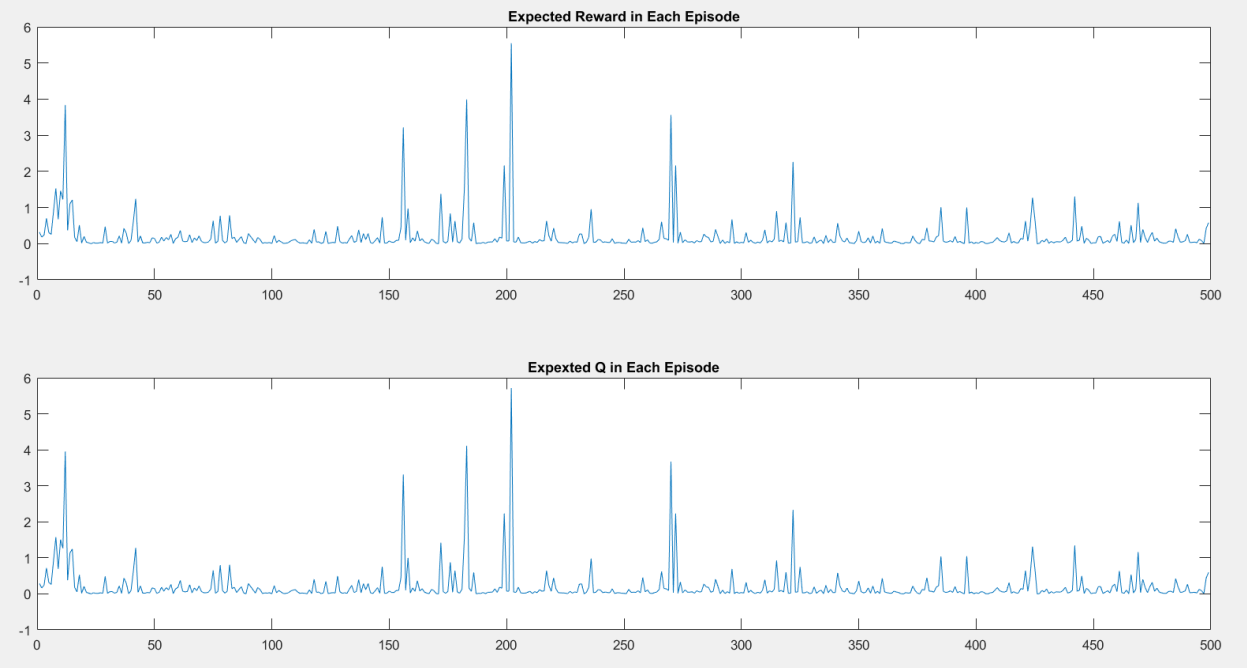
بررسی اثر discount factor:

گاما: 0.02:



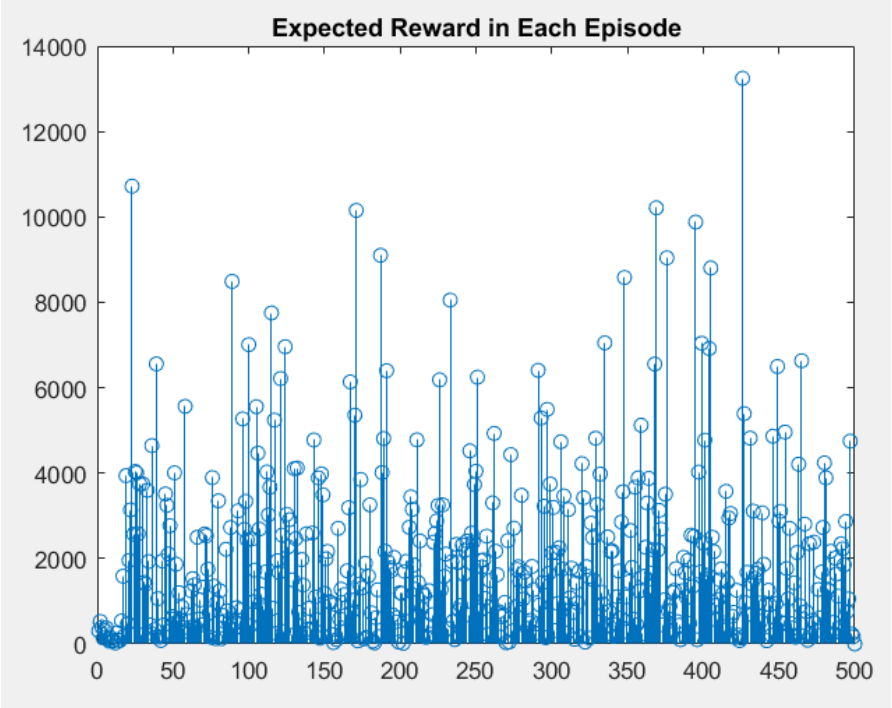
شکل 18- متوسط امتیاز در هر مرحله برای گاما ۰/۰۲

اگر گاما را همچنین کوچکتر از این مقدار قرار دهیم خواهیم داشت که:



شکل 19- متوسط امتیاز در هر مرحله برای گاما ۰/۰۱۵

تعداد مراحل جابه جایی ایجنت در هر اپیزود برابر است با:



شکل 20- تعداد ایتریشن موجو در هر اپیزود

با مشاهده شکل 20 دلیل کم شدن امتیاز جمع اوری شده توسط agent به خوبی قابل مشاهده میباشد. همانطور که میبینید تعداد ایتریشن های در هر ا\یزود بسیار زیاد است به معنی انکه agent دید خوبی از اینده و به تبیت از ان دید خوبی نسبت به محیط اطراف ندارد در نتیجه با حرمت در جهت های غیر صحیح به استیت نهایی خواهد رسید.

با توجه به شکل ها و توضیحات بالا خواهیم داشت:

نرخ تخفیف به ان معنی است که چه مقدار وزنی را به امتیاز های دریافت شده در اینده اختصاص بدهیم. به طور مثال نرخ تخفیف 0 باعث وابستگی مقدار Q به تنهاامتیاز انی و نه اینده خواهد شد و همچنین گامای بزرگ برابر خواهد بود با اثر دادن اینده در اپدیت کردن خروجی.

توجه شود که هر چه این نرخ بزرگ تر باشد در نتیجه آن ِمقدار امتیاز state نهایی که مقدار بزرگی است بیشتر شده و زودتر ایجنت مسیر خود به سمت این state نهایی را خواهد یافت.

همچنین داریم که زمان همگرایی نیز با مقدار گاما رابطه دارد به این شکل که اگر گاما کوچک باشد ِagent باید زمانی زیاد تری را صرف گشت و گذار نماید.

سوال 4:

تعداد دیسک	تعداد مراحل بهینه ترین حالت
3	7
4	15
5	31

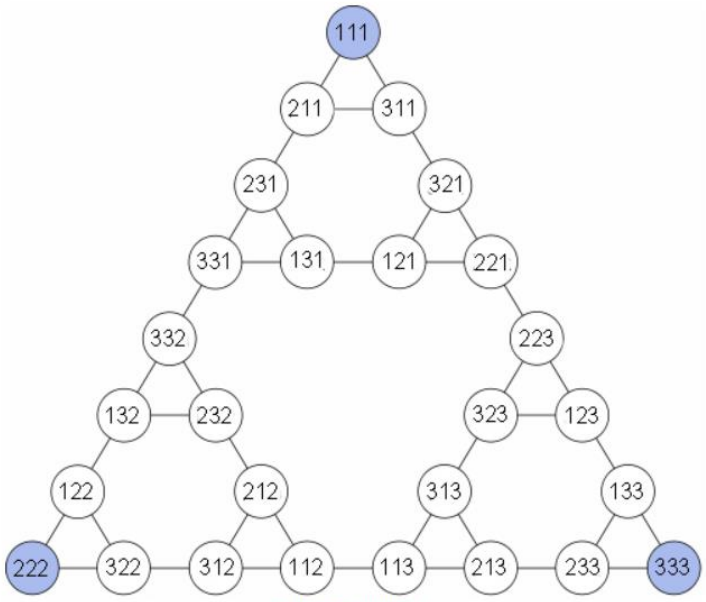
همان طور که مشخص است بهینه ترین تعداد مراحل برای حالت n دیسک برابر است با:

$$2^n-1$$

فرمول بالا یه صورت ریاضی هم اثبات شده است.

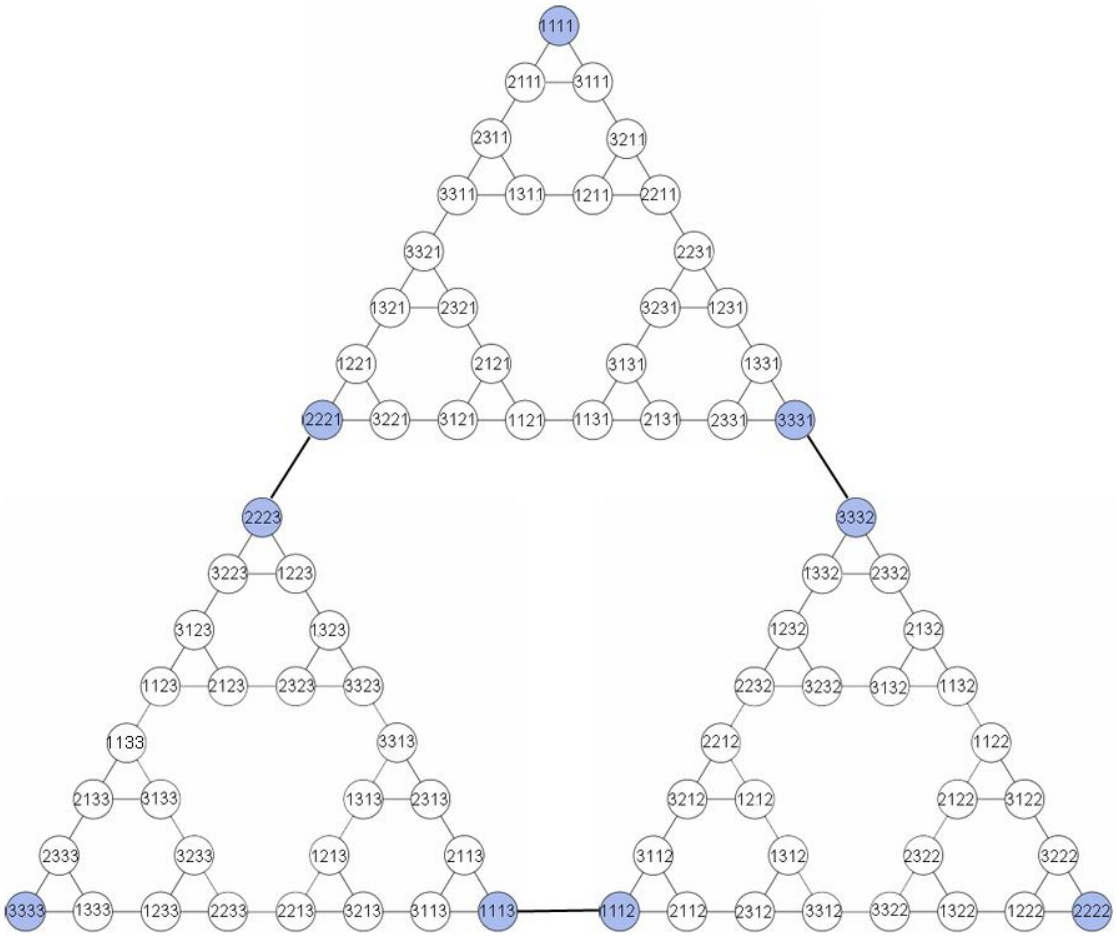
این مسعله را به صورت زیر نیط میتوانید تحقیق کرد:

برای 3 دیسک:



The Hanoi graph H_3.

برای 4 دیسک:


1111
2111 3111
2311 3211
3311 1311 1211 2211
3321 2231
1321 2321 3231 1231
1221 2121 3131 1331
2221 3221 3121 1121 1131 2131 2331 3331
2223 3332
3223 1223 1332 2332
3123 1323 1232 2132
1123 2123 2323 3323 2232 3232 3132 1132
1133 3313 2212 1122
2133 3133 1313 2313 3212 1212 2122 3122
2333 3233 1213 2113 3112 1312 2322 3222
3333 1333 1233 2233 2213 3213 3113 1113 1112 2112 2312 3312 3322 1322 1222 2222

همان طور که مشخص است , این مثلث با اضافه شدن هر دیسک, تکرار خواهد شد.

همچنین کوتاه ترین مسیر برای رسیدن به جواب مسعله پیمودن یک ضلع این مثلث های تکراری میباشد. از انجا که این طول با رابطه 2^n-1 بدست خواهد امد, پس حداقل تعداد مراحل حل مسعله برابر با این مقدار است.

سوال ۵:

در هنگامی برابر بودن مقدار Q value ها ساده ترین راه انتخاب به صورت رندم از بین گزینه های برابر میباشد. به این صورت اگرچه به صورت اتفاقی در یک قدم از قدم های داخل یک اپیزود به انتخاب state بعد پرداخته ایم ِاما از ان جا که تعداد ایتریشن ها یا همان قدم ها بسیار زیاد است ِ این رندم انتخاب کردن باعث واگرایی نخواهد شد و جواب در نهایت همگرا خواهد شد.

راه حل دیگر ان است که از بین موارد دارای مقادیر مساوی ِ state ای به عنوان مرحله بعد انتخاب شود که دارای کمترین نرخ ورود به ان توسط agent میباشد که این مورد به همگرایی نیزکمک خواهد کرد.