به نام او

# درس یادگیری تقویتی

پروژه اول، پوریا صفائی ۹۸۱۱۰۴۰۲

## سوال ۱

تابع generate random samples که در فایل کد قابل مشاهده است وظیفه تولید داده های تصادفی نرمال با امید ریاضی $q_*(a)$ و واریانس ۱ را دارد.(تمام کدهای مربوط در فایل $project۱.py$ در فایل زیپ قرار داده شده است.)

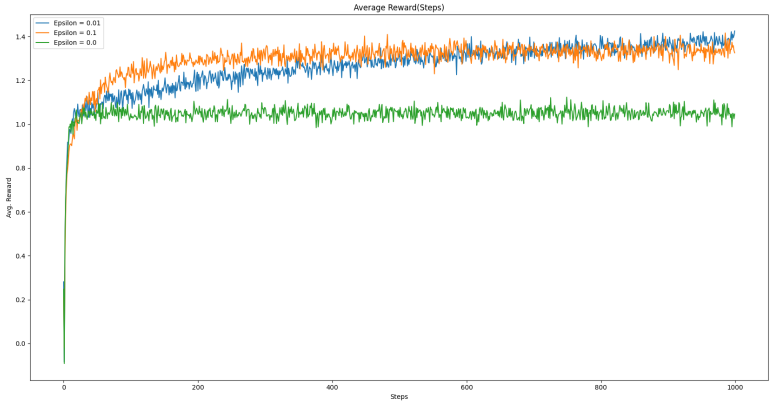
## سوال ۲

با توجه به شکل کشیده شده در کتاب، مقادیر تقریبی هر یک از $q_*(a)$ها را به صورت زیر مینویسیم:

$$q_*(۱) = ۰/۲۵, q_*(۲) = -۰/۶, q_*(۳) = ۱/۵, q_*(۴) = ۰/۵, q_*(۵) = ۱/۲۵,$$
$$q_*(۶) = -۱/۵, q_*(۷) = -۰/۲, q_*(۸) = -۱/۰, q_*(۹) = ۰/۹, q_*(۱۰) = -۰/۵$$

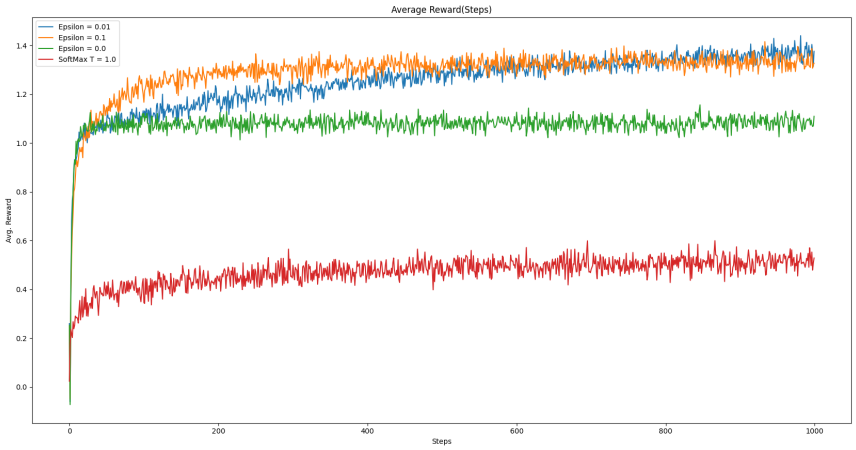
## سوال ۳

همانطور که در نمودار زیر میتوان مشاهده کرد، با فرض اجرای ۱۰۰۰ گام و در هر گام اجرای ۲۰۰۰ مرحله حاصل الگوریتم اپسیلون-حریصانه برای مقادیر مختلف اپسیلون قابل مشاهده است(مقدار اپسیلون صفر معادل همان الگوریتم حریصانه ساده است). همانطور که میتوان این تصویر را با عکس کتاب مشاهده کرد، شباهت کامل مشهود است. به ازای اپسیلون کوچک، پاداش متوسط با شیب کمی در حال افزایش است و به دلیل اینکه اکسپلوریشن کمتر از حالت اپسیلون بزرگتر است، نقطه بهینه خود را کمی دیرتر پیدا میکند. اما چون در این حالت اکسپویتیشن بیشتر است، پس نمودار با یک شیب بزرگتر از حالت دیگر در حال افزایش میانگین پاداش است. حالت اپسیلون صفر نیز با پیدا کردن اولین اکشن که بزرگتر از صفر پاداش دارد، تا انتهای الگوریتم همان عمل را انجام میدهد.



## سوال ۴

در این بخش نیز میتوان نتیجه سافت مکس را برای مقدار دمای T = ۱ مشاهده کرد. همانطور که میبینیم، میزان متوسط پاداش در این حالت نسبت به دیگر حالت ها به طور متوسط کمتر به دست آمده، چرا که دلیل آن اسن است که سافت مکس به اندازه الگوریتم های قبلی حریصانه نیست و اجازه اکسپلوریشن را به ایجنت میدهد. از طرف دیگر چون متوسط توزیع های مختل نزدیک به هم است، پس توزیع احتمالاتی ای که سافتمکس برای حرکت بعدی در نظر میگیرد نیز یکنواخت تر از الگوریتم های حریصانه قبلی است. اما از طرف دیگر میتوان مشاهده کرد که نمودار الگوریتم سافتمکس میتواند صعودی باشد و اگر در مرور زمان اگر مقدار T را کاهش دهیم، این الگوریتم پس از مدتی اکتشاف در ابتدا نهایتا به بهترین جواب میل خواهد کرد.(در اینجا چون مقدار T در مرور زمان ثابت قرار داده شده است، سافتمکس ابتدا امتوسط ارزش همه عمل ها را پیدا میکند و سپس پس از آن با یک توزیع متناسب با پاداش هر عمل، عمل بعدی خود را انتخاب میکند.)




اما همانطور که در نمودار زیر میبینیم(نمودار بنفش)، اگر مقدار T را به مرور زمان کاهش دهیم(در اینجا ما دما را با ضریب ۰/۰۰۵ در هر مرحله کاهش میدهیم)، مشاهده میکنیم که الگوریتم سافتمکس نیز در اینجا به بهینه مقدار ممکن میل میکند.