2019

# گزارش تکلیف شماره ۴

یادگیری ماشین

سید محمد حسینی - ۸۱۰۱۹۴۵۴۱

# سوال ۲:

با توجه به نوشته ی کتاب در صفحه ی ۱۱۵ و ۱۱۶ حداکثر خطا برای تخمین expected n-step return تحت سیاست π از رابطه زیر بدست میآید:

$$\max_s \left| \mathbb{E}_\pi[G_{t:t+n} | S_t = s] - v_\pi(s) \right| \;\leq\; \gamma^n \max_s \left| V_{t+n-1}(s) - v_\pi(s) \right|$$

حال با توجه به اینکه عامل یادگیری از مرحله خاصی شروع میکند هر سری پس شروع اپیزود ها همیشه ثابت است. پس امیدریاضی پاداش ها همیشه از بالا نسبت به ارزش مرحله $S_0$ باند شده است و خطا حداکثر به مقدار سمت راست میشود.

# سوال ۳:

برای یادگیری با توجه به اینکه احتمال جابجایی هارا میدانیم، میتوان از یک off-policy استفاده کرد. به این صورت که با یک سیاست زندگی کنیم و یک سیاست دیگر را evaluate کنیم. برای همین موضوع باید حتما به میزان مناسب exploration انجام شود.

با استفاده از اینکه احتمال را میدانیم، میتوان سیاست را با سرعت بیشتری همگرا کنیم. با دانستن اینکه احتمال جابجایی ها چقدر است، در الگوریتم های حریصانه باعث میشود که بتوان پارامتر ها را راحتتر همگرا کرد و اینکه سیاست ما با استفاده از داده های بیشتر راحت تر تصمیم گیری را انجام دهد.

# سوال ۱:

برای این سوال، با یک gridworld روبرو هستیم. مدل محیط به این شکل خواهد بود که هر حرکت یک پاداش منفی دارند، همچنین در صورت رصد شدن در مقابل دوربین ها و یا دیده شدن توسط نگهبان زندان پاداش منفی بزرگی خواهد داشت. رسیدن به کلید یک پاداش بزرگ دارد و با شرط داشتن کلید، رسیدن به زندانی دوم پاداش بزرگ دیگری خواهد داشت. برای شبیه سازی مدل محیط، یم آرایه ۲ بعدی در نظر میگیریم که در آن نقاط مختلف میزان پاداش و تنبیه مشخص شده باشد. نکته ای که وجود دارد، وجود زندان بان است که باید در هر بار

ران شدن سیستم با احتمال های یکسان، خانه های مشخص شده را تنبیه بزرگی برایش قرار دهیم. برای همین مدل محیط در هر بار ران شدن برای آن محل عوض میشود.

- متاسفانه نتوانستم کد شبیه سازی را انجام دهم.