دانشکده مهندسی برق

یادگیری تقویتی در کنترل

تمرین کلاسی سوم: بررسی روش

Optimistic Initial Value

استاد: دکتر سعید شمقدری

دانشجو: سیده ستاره خسروی

پائیز ۱۴۰۳

چکیده

در این تمرین به تحلیل و پاسخ تمرین کلاسی سری سوم پرداخته می‌شود.

واژه‌های کلیدی: یادگیری تقویتی، راهزن چند دست

فهرست مطالب

عنوان صفحه

فهرست تصاویر و نمودارها

عنوان صفحه

فصل ۱: شبیه‌سازی مسئله راهزن چنددست

۱.۱ صورت سوال

در گراف مقایسه روش‌های مختلف حل مسئله MAB در کتاب (شکل ۶-۲) بررسی کنید در روش Optimistic، چرا برایQ_0=1 ، عملکرد از سایر مقادیر بهتر است؟

۱.۲ پاسخ تحلیلی و شبیه سازی

ابتدا برای پاسخ به این سوال لازم است، به نمودار نگاهی بیندازیم.

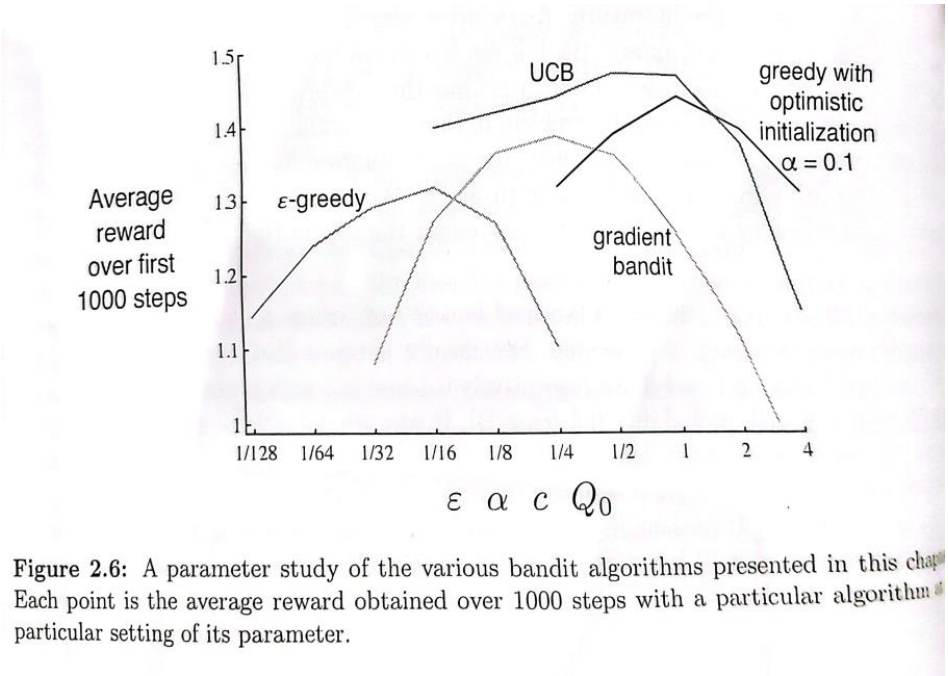


Figure 2.6: A parameter study of the various bandit algorithms presented in this chap Each point is the average reward obtained over 1000 steps with a particular algorithm particular setting of its parameter.

شکل ۱-۱: نمودار مقایسه‌ای روش‌های حل مسئله راهزن چنددست

گراف مربوط به روش شرایط اولیه خوشبینانه، بیانگر این است که در این روش، مسئله راهزن چنددست با روش حریصانه، گام ۰.۱ و شرایط اولیه‌ی خوشبینانه برابر با ۱ بیشترین پاداش میانگین را در ۱۰۰۰ گام اولیه به ارمغان می‌آورد. این در حالی است که بحث شد هره این مقدار اولیه‌ی خوشبینانه بیشتر باشد بهتر است ولی در این نمودار چنین چیزی مشاهده نمی‌گردد.

باید به معادله‌ی زیر توجه کنیم.

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n) = Q_n + 0.1(R_n - Q_n) = 0.9Q_n + 0.1R_n$$

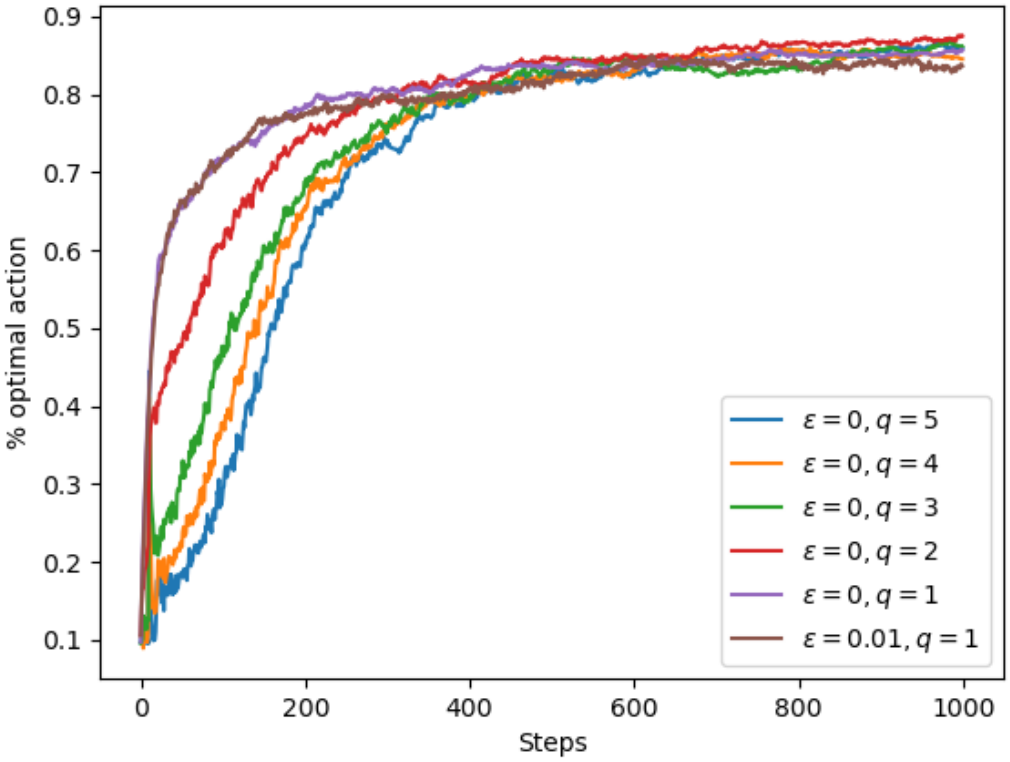
باید توجه نمود که هرچقدر مقدار Q_0 بیشتر باشد، اثر آن دیرتر از بین می‌رود و برای ازبین رفتن اثر آن زمان بیشتری لازم است. با توجه به اینکه در اکتشاف عملا یادگیری رخ نمی‌دهد، و انتخاب این مقدار اولیه با مقداری بیش از حد بالا، عامل ممکن است برای مدت طولانی‌تری به جست و جو ادامه دهد، و اصلاح این براورد اولیه زمان بسیاری طول می‌کشد. این امر باعث تأخیر در رسیدن به بهترین عمل می‌گردد و ممکن است در مراحل اولیه عامل به عملکردی زیربهینه برسد. ظاهرا تنظیم مقدار ۱ برای مقدار اولیه خوشبینانه، تعادل مناسبی میان اکتشاف و بهره برداری ایجاد می‌کند، به اندازه کافی برای انجام جست و جو خوشبینانه است، ولی انقدر هم زیاد نیست که عامل وقت زیادی را برای جست و جو صرف کند. برای مقادیر بالاتر از ۱، عامل ممکن است بیش از حد نیاز به جست و جو بپردازد.

در حالی که مقادیر اولیه بالا مثل ۵، می‌توانند برای جست و جو مفید باشند، تنظیم آنها با مقداربیش از حد بالا می‌تواند باعث همگرایی کندتر و عملکرد ضعیف‌تر شود، زیرا عامل برای مدت طولانی به بررسی و جست و جو ادامه می‌دهد. مقدار اولیه ۱ در این مورد به نظر می‌رسد که یک تعادل مناسب بین جست و جو و بهره‌برداری را برقرار می‌کند.

برای بررسی بیشتر این موضوع، با استفاده از کدی که برای تمرین سری دوم قسمت UCB نوشته شده بود، شبیه سازی حریصانه با مقدار اولیه خوشبینانه را برای ۵ مقدار متفاوت انجام دادیم.

در شبیه سازی مقدار α برابر با ۰.۱ لحاظ کردیم، شرایط اولیه خوشبینانه نیز از ۱ تا ۵ لحاظ شدند، روش نیز greedy است و به تبع از آن مقدار ε برابر با صفر است. در یک حالت نیز با مقدار ε برابر با ۰.۰۱ نیز شبیه سازی را انجام دادیم که محل بحث نیست.

خروجی شبیه سازی در نمودار شکل ۲-۱ آورده شده است.



شکل ۲-۱: خروجی شبیه‌سازی

در نمودار فوق مشاهده می‌گردد که با افزایش مقدار اولیه خوشبینانه، مدت زمانی که عامل صرف جست و جو می‌کند و به خاطر آن عمل‌های غیربهینه را انتخاب می‌کند و به تبع از آن پاداش کمتری دریافت می‌کند، طولانی‌تر می‌شود، در حالاتی که مقدار اولیه‌ی خوشبینانه بیشتر از ۱ است، فرایند یادگیری و همگرایی کندتر می‌شود. مطابق توضیحات ارائه شده در صفحه قبل در اجرای با مقدار اولیه خوشبینانه ۱ تعادل بهتری میان اکتشاف و بهره برداری برقرار می‌گردد. به دلیل کند شدن روند همگرایی در مقادیر بالاتر از ۱ و اینکه نمودار ۱-۱ نیز مربوط به ۱۰۰۰ اجرای اولیه است، میانگین پاداش کسب شده با آن مقادیر کمتر از حالتی است که مقدار اولیه‌ی ۱ تنظیم می‌شود. درصورت ادامه یافتن فرایند شبیه‌سازی انتظار می‌رود که وضعیت حالاتی که مقدار اولیه بیشتر از ۱ است بهتر شود. این موضوع از نمودار ۲-۱ نیز قابل استنتاج است، درواقع وقتی مقدار اولیه را افزایش می‌دهیم باید فرصت بیشتری نیز به عامل بدهیم تا اثر این مقدار اولیه نیز از بین برود.

۱.۳ منابع

کتاب Sutton and Barto

ریپازیتوری‌های:

https://github.com/setarekhosravi/reinforcement-learning-an-introduction