گزارش اجرای الگوریتم Q-Learning روی محیط Lake Frozen

سونیا سیف ۴۰۰۱۰۰۲۷۶

۱۳ تیر ۱۴۰۴

علوم اعصاب، دانشکده برق دانشگاه صنعتی شریف
استاد: دکتر حمید آقاجان

۱ مروری بر مفاهیم اولیه:

۱.۱ تساوی بلمن و نقش γ و α

تابع ارزش بهینه Q^* تحت تساوی بلمن به صورت زیر تعریف می‌شود:

$$Q^*(s,a) \;=\; \sum_{s'} P(s' \mid s,a)\left[R(s,a,s') \;+\; \gamma \max_{a'} Q^*(s',a')\right]$$

در این رابطه:

- $P(s' \mid s,a)$ احتمال انتقال به حالت s' پس از انجام عمل a در حالت s است.
- $R(s,a,s')$ پاداش دریافت‌شده در این انتقال است.
- $\gamma \in [0,1]$ (discount factor) تأثیر پاداش‌های آینده را تعیین می‌کند:
  - $\gamma = 0$: توجه صرفاً به پاداش فوری.
  - $\gamma \to 1$: توجه بیشتر به پاداش‌های بلندمدت.

در Q Learning ، به‌جای مجموع احتمالات، از نمونه‌برداری استفاده می‌کنیم و مقدار $Q(s_t,a_t)$ را به‌صورت زیر به‌روزرسانی می‌کنیم:

$$Q(s_t,a_t) \;\leftarrow\; Q(s_t,a_t) \;+\; \alpha\left[r_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\right]$$

که در آن $\alpha \in (0,1]$ (rate learning) سرعت به‌روزرسانی را کنترل می‌کند:

- α بزرگ: تغییرات سریع اما نوسان بیش‌تر.
- α کوچک: همگرایی پایدار اما کند.

محیط‌های قطعی ($P(s' \mid s,a) \in \{0,1\}$) همگرایی سریع‌تر ولی محیط‌های تصادفی ،با احتمال کمتر از یک، نیازمند تعداد اپیزود و اکتشاف بیشتری هستند.

۲.۱ سیاست ϵ- greedy و نقش ϵ

برای حفظ تعادل میان اکتشاف (exploration) و بهره‌برداری (exploitation) ، در هر گام از سیاست زیر استفاده می‌شود:

$$a_t = \begin{cases} \text{عمل تصادفی} & \text{با احتمال } \epsilon \\ \arg\max_a Q(s_t,a) & \text{با احتمال } 1-\epsilon \end{cases}$$

در ابتدای آموزش $\epsilon = \epsilon_{\max} = 1.0$ است تا عامل فضای حالت-عمل را بکاود. سپس برای تمرکز بر بهره‌برداری، ϵ را با معادله‌ی زیر به مرور کاهش می‌دهیم:

$$\epsilon_{\text{new}} = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})\, e^{-\lambda \times \text{episode}},$$

که معمولاً $\epsilon_{\min} = 0.01$، $\lambda \approx 0.01$. اگر ϵ خیلی سریع کاهش یابد، عامل ممکن است پیش از کشف بهترین سیاست به دام سیاست‌های محلی بیفتد، و اگر خیلی کند کاهش یابد، زمان زیادی صرف اکتشاف می‌شود و همگرایی کند می‌گردد. تنظیم مناسب λ کلید دستیابی به نرخ موفقیت بالا است.

۲ شرایط دستیابی به استراتژی بهینه

با توجه به مفاهیم اولیه ذکر شده، در زیر جمعبندی شرایط لازم برای رسیدن به استراتژی بهینه را آورده ام:

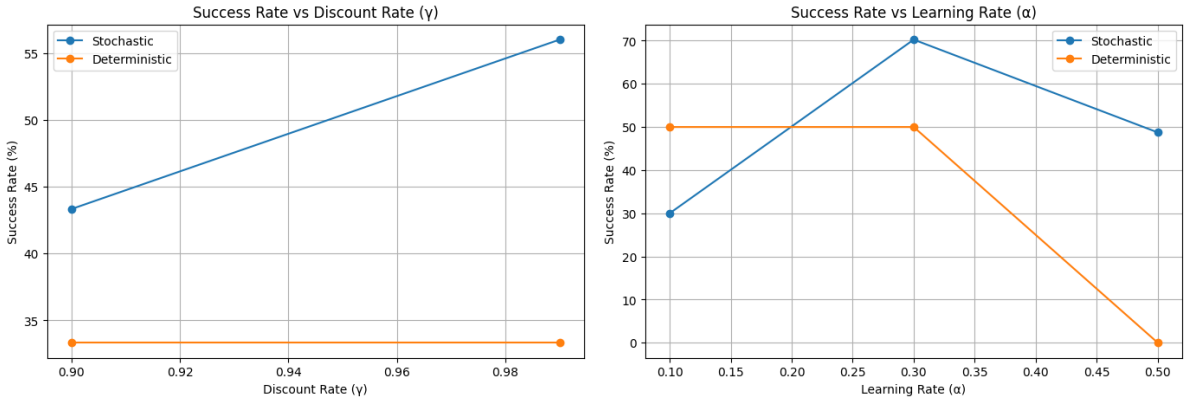
محیط قطعی (Deterministic)

- هر عمل یک پیامد مشخص دارد ($P(s' \mid s, a) = 1$) که همگرایی سریع و بدون نویز را ممکن می‌سازد.
- $\gamma \approx 0.99$: توجه بلندمدت به پاداش‌ها تضمین می‌کند مسیرهای دوردستی با پاداش کلان شناسایی شوند.
- $\alpha \approx 0.1$: سرعت یادگیری متعادل، هم سریع و هم پایدار.
- ϵ- greedy با decay نمایی مناسب ($\epsilon_{\max} = 1 \to \epsilon_{\min} = 0.01$) و تعداد کافی اپیزود (مثلاً ۵۰۰۰).

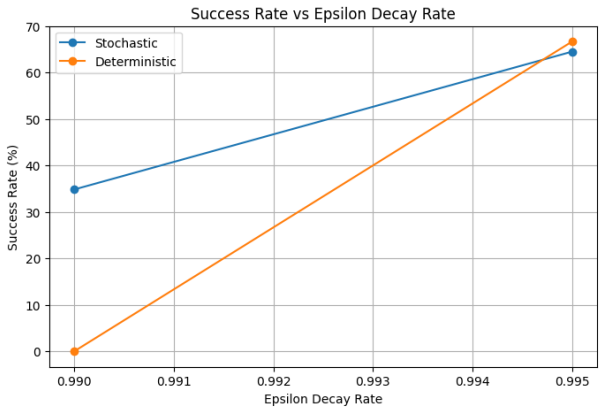
محیط تصادفی (Stochastic)

- انتقال‌ها با احتمالات مختلف رخ می‌دهند ($P(s' \mid s, a) < 1$) و نیازمند داده‌های بیشتر برای میانگین‌گیری است.
- γ و α مشابه محیط قطعی، اما α گاهی باید کوچکتر مثلاً 0.05، 0.1تا نوسان کاهش یابد.
- decay آهسته‌تر ϵ برای حفظ اکتشاف بیشتر در اپیزودهای اولیه.
- اپیزودهای بیشتر (مثلاً ۱۰۰۰۰) تا خطای بلمن در میانگین نمونه‌ها همگرا گردد.

پس انتظار داریم در هر دو حالت،با تنظیم دقیق γ، α، و روند کاهش ϵ همراه با تعداد مناسب اپیزودها باعث افزایش موفقیت عامل شود. حال نتایج بدست آمده از کد را بررسی میکنیم تا ببینیم با آنچه انتظار داریم همخوانی دارد یا خیر.



(آ) (ب)



(ج)

شکل ۱

۳ بررسی نمودارها

با توجه به انتظاراتی که مطرح شد، نتایج به‌دست‌آمده را به‌صورت زیر می‌توان تحلیل کرد:

تأثیر γ (فاکتور تخفیف)

- در محیط تصادفی ،(Stochastic) با افزایش γ از 0.90 به 0.99، نرخ موفقیت از 43% به 56% افزایش یافته است. این درست است چون با γ بالاتر عامل پاداش‌های دوردست را جدی‌تر می‌گیرد و مسیر بهینه‌تر را یاد می‌گیرد.
- در محیط قطعی ،(Deterministic) تغییر γ اثری بر نرخ موفقیت ($\approx 33\%$) نداشته است. این نشان می‌دهد که در این آزمایش احتمالاً سیاست بهینه برای هر γ در عمل یکسان بوده یا تعداد اپیزود برای همگرایی کافی نبوده است.

تأثیر α (نرخ یادگیری)

- در محیط تصادفی نرخ موفقیت برای $\alpha = 0.10, 0.30, 0.50$ به ترتیب $\approx 30\%$، $\approx 70\%$، $\approx 49\%$ است. به‌خوبی مشاهده می‌شود که مقدار میانه ($\alpha = 0.30$) تعادل بهتری میان سرعت یادگیری و پایداری ایجاد کرده است.
- در محیط قطعی نرخ موفقیت برای $\alpha = 0.10, 0.30$ حدود 50% بوده و برای $\alpha = 0.50$ به 0% رسیده است. این رفتار نشان می‌دهد که α خیلی بزرگ باعث نوسان بیش از حد در Q-Table و در نهایت یادگیری ناموفق می‌شود.

تأثیر ϵ

- افزایش آن موجب افزایش موفقیت عامل شده است. که با مفاهیم قبلی همخوانی دارد؛ به خصوص در محیط های کوچک.
- همچین تاثیر افزایش اپسیلون بر موفقیت عامل در محیط قطعی بسیار چشگیر تر از محیط تصادفی است که آن نیز مورد انتظار ما بود.

۴ عملکرد عامل، نقدی بر این روش آموزش عامل

با توجه به مقادیری که برای هایپرپارامترهای عامل بدست آوردیم؛ به سراغ آموزش عامل رفتم. و مشکلی که به هنگام آموزش به آن برخوردم و راه حل هایی کمک کننده را در ادامه آورده ام. در واقع میتوان این قسمت را ایراد و نقدی به این روش آموزش عامل دانست. همچنین به این علت نمیدانم آهنگ موفقیت عامل را چقدر باید گزارش دهم.

تا زمانی که عامل در محیط قطعی به‌طور تصادفی مسیر درست را نبیند، همهٔ مقادیر Q برابر صفر می‌مانند و سیاست ϵ-greedy عملاً تصادفی خواهد بود. برای اینکه قبل از اولین رسیدن به هدف، شانس عامل را برای کشف سریع‌تر مسیر بهبود دهیم، پیشنهادات زیر ارائه میشود: (این روش ها از سرچ بدست آمده و تست نشده اند.)

۱. مقداردهی اولیهٔ خوش‌بینانه (Optimistic Initialization)
   - مقداردهی اولیهٔ $\mathbf{Q}(s,a) = Q_0 > 0$ (مثلاً $Q_0 = 1$)
   - تمام اعمال در نگاه اول جذاب به نظر می‌رسند و عامل آن‌ها را امتحان می‌کند تا مقدار واقعی‌شان را بیابد.

۲. شکل‌دهی پاداش (Reward Shaping)
   - تعریف پاداش پادنامنفی کوچک برای نزدیک‌شدن به هدف (مثلاً کاهش فاصله منانه).
   - پاداش اضافی برای هر گامی که عامل را به خانه هدف نزدیک‌تر می‌کند.

۳. نمایش رفتار نمونه (Learning Imitation)
   - ارائه یک مجموعه اپیزودهای نمونهٔ درست (trajectory) به عامل در ابتدا.

۴. شروع تصادفی از نقاط میانی
   - در برخی اپیزودها، عامل را نزدیک به هدف یا وسط مسیر قرار دهیم.
   - به مرور فاصلهٔ شروع را افزایش دهیم تا عامل همهٔ مسیر را یاد بگیرد.

با اعمال این روش‌ها، حتی قبل از اولین برخورد با خانهٔ هدف، Q-Table مقداری غیر صفر خواهد گرفت و عامل شانس بیشتری برای یافتن مسیر بهینه در اولین اجرای کامل خواهد داشت. این کار باعث می‌شود اولین موفقیت سریع‌تر نمایان شود.

۵ ضبط GIF اجرای عامل

در این بخش، یک تصویر متحرک (GIF) از رفتار عامل در محیط قطعی ارائه شده است. برای تولید این GIF، پس از اولین باری که عامل موفق به رسیدن به خانه هدف شد، مراحل اجرا به‌صورت کامل ضبط شد تا سیاست بهینه و استقرار رفتار عامل پس از یادگیری آشکار شود. در نتیجه، GIF نمایش‌دهندهٔ اجرای عامل پس از دریافت اولین پاداش مثبت و استفاده‌ی خالص از سیاست greedy است.

۶ دسترسی به کد

کد کامل مربوط به این پروژه را می‌توانید در آدرس زیر مشاهده و دانلود کنید:

```
https://colab.research.google.com/drive/
1J3tp7XrSYrEvqduP-aIuBrNDoCnttZdP?usp=sharing>
```

همچنین گیف مربوطه و فایل کد نیز در داخل همین پوشه قابل دسترسی است.