دانشکده مهندسی برق

عنوان:

گزارش نهایی پروژه درس یادگیری تقویتی

حل هزارتوی پویا با **Value Iteration**، **Q-Learning**، **SARSA(λ)** و انتقال یادگیری

مینا امیرخانی

۴۰۴۰۳۳۸۴

استاد درس:

دکتر خواسته

مرداد ماه ۱۴۰۵

شماره دانشجویی	40403384
پایه Seed	8
اندازه هزارتو	15 × 15
مخزن پروژه	https://github.com/mina2collab/RL_FinalProject_40403384

فهرست بخش‌ها

- ۱. چکیده
- ۲. تعریف مسئله و محیط
- ۳. الگوریتم‌ها و تنظیمات آزمایش
- 4.نتایج Value Iteration
- ۵. نتایج Q-Learning
- ۶. نتایج SARSA(λ)
- ۷. مقایسه سیاست‌ها و تحلیل اختلاف
- ۸. انتقال یادگیری
- ۹. رابط گرافیکی، آزمون‌ها و بازتولیدپذیری
- ۱۰. محدودیت‌ها و پیشنهادهای بهبود
- ۱۱. جمع بندی و نتیجه گیری
- ۱۲. استفاده از ابزار هوش مصنوعی

۱. چکیده

در این پروژه یک محیط هزارتوی پویا، تصادفی و قابل بازتولید طراحی شد و سه روش اصلی یادگیری تقویتی شامل Value Iteration، Q-Learning و SARSA(λ) روی یک تعریف حالت و پاداش مشترک بررسی شدند. حالت محیط به صورت (row, column, has_key, gate_phase) تعریف شد تا وضعیت کلید و فاز دروازه دوره‌ای در مدل مارکوف لحاظ شود. در ادامه، انتقال جدول Q از نقشه مبدأ به دو نقشه مقصد مشابه و متفاوت با روش‌های انتقال کامل، مقیاس‌شده و انتخابی مطالعه شد. خروجی‌های عددی به صورت CSV و JSON، مدل‌ها به صورت فایل JSON و نمودارها به صورت PNG ذخیره شده‌اند. نتایج نشان می‌دهد انتقال دانش در محیط مشابه می‌تواند عملکرد اولیه را به‌طور محسوسی افزایش دهد، در حالی که در محیط متفاوت احتمال انتقال منفی وجود دارد و انتقال انتخابی می‌تواند ریسک آن را کاهش دهد.

۲. تعریف مسئله و محیط

Seed پایه از رقم ماقبل آخر شماره دانشجویی استخراج شده است؛ بنابراین Seed برابر ۸ و اندازه نقشه برابر ۱۵ + (8 mod 4) = 15 است. نقشه شامل دیوار، خانه عادی، خانه جریمه، نقطه شروع، کلید، در قفل، هدف و دروازه دوره‌ای است. عامل باید ابتدا کلید را جمع‌آوری کند، از در عبور کند و سپس به هدف برسد.

۲/۱. فضای حالت و عمل

$$\text{state} = (\text{row}, \text{column}, \text{has_key}, \text{gate_phase})$$

- row و column موقعیت عامل را مشخص می‌کنند.
- has_key نشان می‌دهد کلید جمع‌آوری شده است یا نه.
- gate_phase فاز فعلی دروازه دوره‌ای را ثبت می‌کند و برای حفظ خاصیت مارکوف لازم است.
- اعمال عامل شامل بالا، پایین، چپ و راست هستند.

۲/۲. انتقال تصادفی و پاداش

در هر گام، عمل انتخابی با احتمال ۰.۸ اجرا می‌شود و هر یک از دو عمل عمود بر آن با احتمال ۰.۱ رخ می‌دهد. برخورد با دیوار باعث باقی‌ماندن عامل در همان خانه و دریافت جریمه می‌شود. دو حالت پاداش sparse و shaped پیاده‌سازی شده‌اند. پاداش شکل‌داده‌شده برای کاهش زمان یادگیری، بازخورد میانی فراهم می‌کند؛ با این حال ارزیابی نهایی بر نرخ موفقیت، تعداد گام، کیفیت مسیر و مقایسه سیاست متکی است.

۲/۳. مشخصات کامل MDP و تابع پاداش

نقشه مبدأ دارای ابعاد ۱۵×۱۵ و در مجموع ۲۲۵ خانه است. از این تعداد، ۸۴ خانه دیوار و ۱۴۱ خانه قابل‌عبور هستند. حالت مارکوف به‌صورت state = (row, column, has_key, gate_phase) تعریف شده است. با توجه به دو مقدار ممکن برای has_key و دو فاز دروازه دوره‌ای، فضای حالت شمارش‌شده شامل $141 \times 2 \times 2 = 564$ حالت است. فضای عمل شامل چهار حرکت بالا، پایین، چپ و راست است. دروازه دوره‌ای دارای دوره ۲ است و در فاز صفر باز می‌شود. حداکثر طول هر اپیزود سه برابر تعداد خانه‌های قابل‌عبور، یعنی $3 \times 141 = 423$ گام، در نظر گرفته شده است. اپیزود با رسیدن عامل به هدف پس از دریافت کلید با موفقیت پایان می‌یابد؛ در صورتی که عامل تا سقف ۴۲۳ گام به هدف نرسد، اپیزود به‌صورت ناموفق متوقف می‌شود.

تعداد دیوارها، دوره دروازه و موقعیت اجزای نقشه مستقیماً در فایل نقشه ثبت شده‌اند؛ کد محیط نیز ۵۶۴ حالت را با ترکیب خانه‌های قابل‌عبور، وضعیت کلید و فاز دروازه می‌سازد و سقف اپیزود را سه برابر خانه‌های قابل‌عبور تعیین می‌کند .

۲-۳-۱. تابع پاداش پایه

پاداش نهایی در حالت Sparse	رویداد
-1	حرکت عادی
-5	برخورد با دیوار یا خروج از نقشه
-8	تلاش برای عبور از درِ بسته بدون کلید
-4	تلاش برای عبور از دروازه دوره‌ای بسته
-10	ورود به خانه جریمه
+20	جمع‌آوری کلید
+100	رسیدن به هدف پس از دریافت کلید
-20 علاوه بر پاداش همان گام	رسیدن به سقف گام‌ها

این اعداد دقیق‌اند، چون محیط ابتدا برای هر عمل هزینه -1 در نظر می‌گیرد و سپس جریمه یا پاداش رویداد را به آن اضافه می‌کند؛ مثلاً جریمه خام برخورد با دیوار -4 است، اما پاداش نهایی آن گام برابر -5 می‌شود. در مورد پایان اپیزود نیز -20 به پاداش رویداد گام آخر اضافه می‌شود.

در طراحی تابع پاداش، هزینه ثابت -1 برای هر حرکت از پرسه‌زدن بی‌هدف جلوگیری می‌کند. جریمه برخورد با دیوار، در بسته، دروازه بسته و خانه‌های خطرناک باعث می‌شود عامل مسیرهای ایمن‌تر را ترجیح دهد. پاداش +20 برای کلید، عامل را به تکمیل مرحله میانی مأموریت هدایت می‌کند و پاداش +100 هدف تضمین می‌کند که پایان موفق اپیزود از جمع‌آوری پاداش‌های فرعی ارزش بیشتری داشته باشد. همچنین جریمه اضافی سقف گام‌ها، اپیزودهای طولانی و ناموفق را نامطلوب می‌کند.

۲-۳-۲. پاداش شکل‌داده‌شده

در حالت Shaped ، پاداش پایه Sparse حفظ می‌شود و یک جمله پتانسیلی به آن افزوده می‌شود:

$$F(s,s') = 0.20 \times [\gamma\Phi(s') - \Phi(s)]$$

پیش از جمع‌آوری کلید، تابع پتانسیل بر اساس فاصله منهتن عامل تا کلید محاسبه می‌شود و پس از جمع‌آوری کلید، فاصله تا هدف ملاک قرار می‌گیرد. در نتیجه، حرکت در جهت مقصد موقت بازخورد مثبت‌تری ایجاد می‌کند و حرکت دورشونده بازخورد ضعیف‌تر یا منفی می‌گیرد. این روش بدون حذف پاداش اصلی، اطلاعات میانی بیشتری برای تسریع یادگیری فراهم می‌کند .

۳. الگوریتم‌ها و تنظیمات آزمایش

۳.۱ Value Iteration

Value Iteration با استفاده از مدل کامل انتقال و به‌روزرسانی بلمن پیاده‌سازی شده است. مقادیر gamma برابر ۰.۸۰، ۰.۹۰ و ۰.۹۵ مقایسه شدند. معیارهای اصلی شامل تعداد تکرار، زمان اجرا، delta نهایی و توافق سیاست هستند.

۳.۲ Q-Learning

Q-Learning یک روش model-free و off-policy است. سیاست رفتار epsilon-greedy بوده و دو برنامه کاهش epsilon، یعنی کاهش خطی و نمایی، اجرا شده‌اند. فرمول به‌روزرسانی به صورت زیر است:

$$Q(s,a) \leftarrow Q(s,a) + \text{alpha} * [r + \text{gamma} * \max_a' Q(s',a') - Q(s,a)]$$

۳.۳ SARSA(λ)

SARSA(λ) به صورت on-policy و با ردپای شایستگی replacing پیاده‌سازی شده است. انتخاب replacing باعث می‌شود در بازدید مجدد یک زوج حالت-عمل، مقدار ردپا به جای انباشته‌شدن نامحدود، جایگزین شود و پایداری عددی بهتری ایجاد کند. مقادیر lambda برابر ۰.۰، ۰.۳، ۰.۷ و ۰.۹ بررسی شده‌اند. در lambda=0 روش به SARSA یک‌مرحله‌ای نزدیک می‌شود و با افزایش lambda، خطای TD به حالت‌ها و اعمال گذشته منتقل می‌شود.

۳.٤ تنظیمات مشترک آزمایش‌ها

پارامتر	Value Iteration	Q-Learning	SARSA(λ)
نوع روش	Model-based	Off-policy ،Model-free	On-policy ،Model-free
پاداش اجرای اصلی	Sparse	Shaped	Shaped
ضریب کاهش γ	۰.۹۵ ،۰.۹۰ ،0.80	0.95	0.95
نرخ یادگیری α	ندارد	0.10	0.10
ε شروع	ندارد	1.00	1.00
ε پایان	ندارد	0.05	0.05
برنامه کاهش ε	ندارد	خطی و نمایی	خطی
تعداد اپیزود آموزش	تا همگرایی بلمن	600	600
تعداد اپیزود ارزیابی	۱۰۰	100	100

برای حفظ بازتولیدپذیری، Seed پایه تمام اجراهای اصلی برابر ۸ در نظر گرفته شد. Value Iteration تا رسیدن تغییر مقادیر حالت به آستانه همگرایی اجرا شد، در حالی که Q-Learning و SARSA با تعداد ثابت اپیزود آموزش دیدند. نتایج فصل‌های ۵ و ۶ مربوط به اجرای اصلی با پاداش Shaped هستند؛ مقایسه منصفانه هر سه الگوریتم با پاداش مشترک Sparse به‌صورت جداگانه در فصل ۷ گزارش شده است.

۳-۵. مقایسه روش‌های Model-based و Model-free

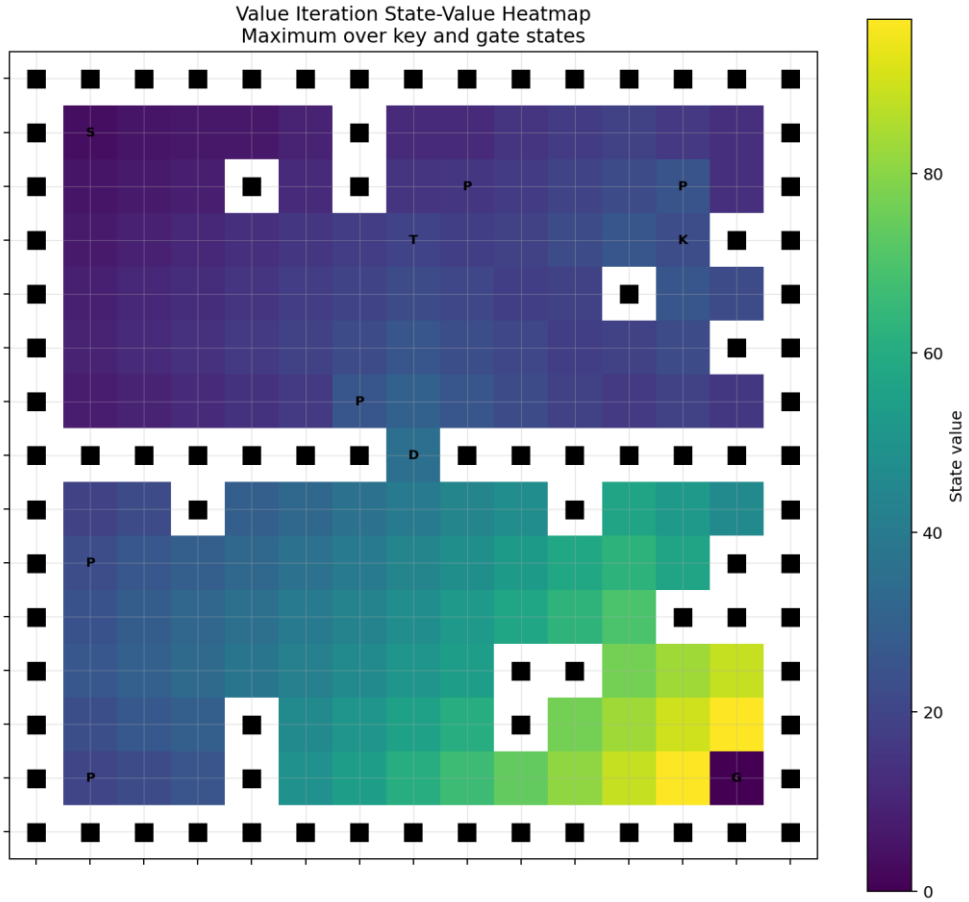
Value Iterationیک روش Model-based است؛ زیرا برای محاسبه مقدار موردانتظار هر عمل به مدل کامل انتقال P(s'|s,a) و تابع پاداش محیط نیاز دارد. مزیت این روش آن است که در صورت معلوم‌بودن مدل و ثابت‌بودن پارامترها، روندی قطعی دارد و می‌تواند سیاست مرجع را بدون تجربه‌کردن اپیزودهای واقعی استخراج کند. در مقابل، ساخت و پیمایش مدل انتقال برای فضاهای حالت بزرگ هزینه محاسباتی و حافظه بیشتری ایجاد می‌کند و در محیط‌های ناشناخته ممکن است مدل دقیق در دسترس نباشد.

Q-Learningو SARSA روش‌های Model-free هستند و تنها از نمونه‌های مشاهده‌شده شامل حالت، عمل، پاداش و حالت بعدی استفاده می‌کنند. این ویژگی آن‌ها را برای محیط‌هایی که مدل انتقال ناشناخته است مناسب می‌سازد؛ بااین‌حال به تعداد زیادی تعامل و اپیزود آموزشی نیاز دارند و نتایج آن‌ها بهSeed ، میزان اکتشاف و تنظیم فراپارامترها حساس است Q-Learning .یک روش Off-policyاست و در هدف TD از بهترین عمل فرضی حالت بعدی استفاده می‌کند، در حالی که SARSA به‌صورت On-policy عمل واقعی انتخاب‌شده توسط سیاست رفتار را وارد به‌روزرسانی می‌کند.

۴. نتایج Value Iteration

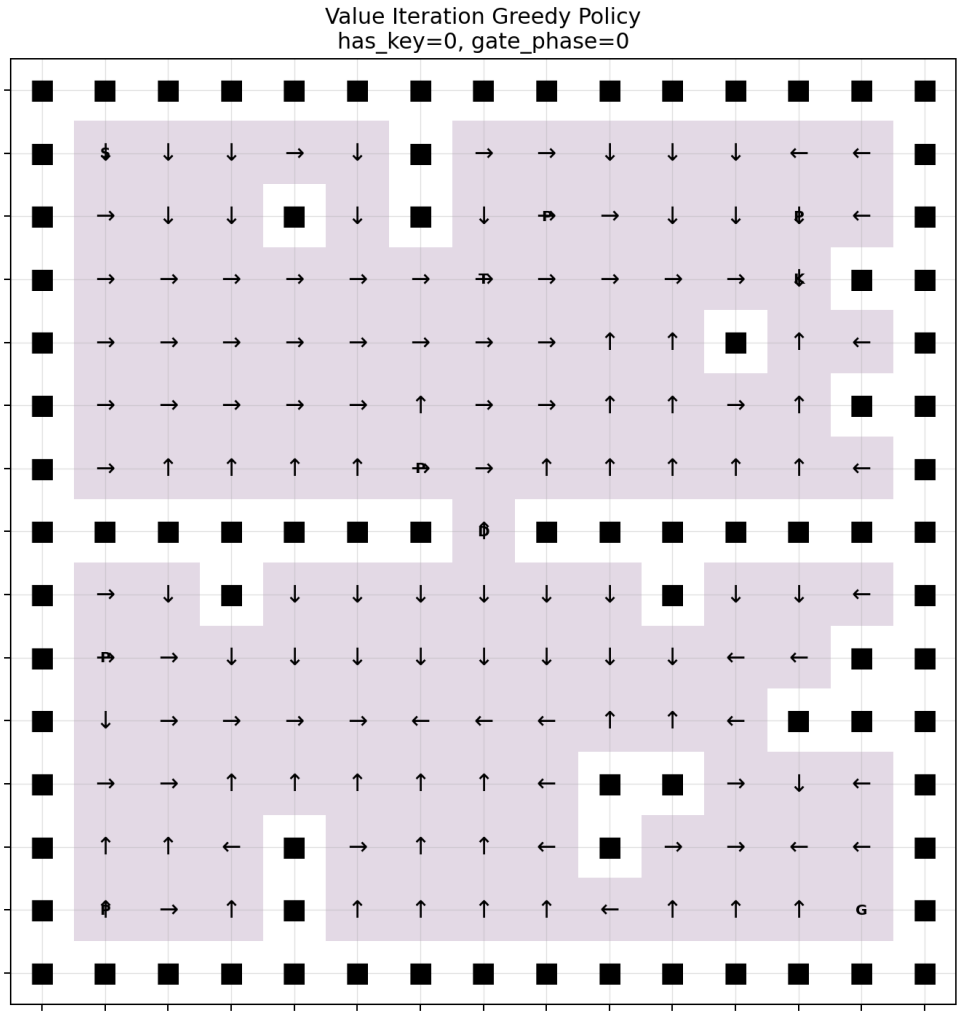
نتایج این بخش مربوط به اجرای Value Iteration روی نقشه مبدأ با پاداش Sparse و Seed پایه ۸ است. الگوریتم با استفاده از مدل کامل انتقال تصادفی محیط اجرا شد؛ در هر حالت و عمل، احتمال اجرای عمل موردنظر برابر ۰.۸ و احتمال اجرای هرکدام از دو عمل عمود برابر ۰.۱ در نظر گرفته شد. مقادیر ضریب کاهش γ برابر ۰.۸۰، ۰.۹۰ و ۰.۹۵ آزمایش شدند و تکرارهای بلمن تا رسیدن بیشینه تغییر مقدار حالت‌ها به آستانه همگرایی ادامه یافتند. سیاست حاصل از γ=0.95 به‌عنوان سیاست مرجع انتخاب شد و درصد توافق دو اجرای دیگر نسبت به همین سیاست محاسبه گردید.

gamma	تکرار	زمان)ثانیه(	نهایی delta	توافق سیاست
0.8	84	1.27	$< 10^{-8}$	91.46%
0.9	176	2.58	$< 10^{-8}$	98.22%
0.95	361	5.20	$< 10^{-8}$	100.00%



شکل ۱ **Heatmap** - مقدار حالت حاصل از **Value Iteration**

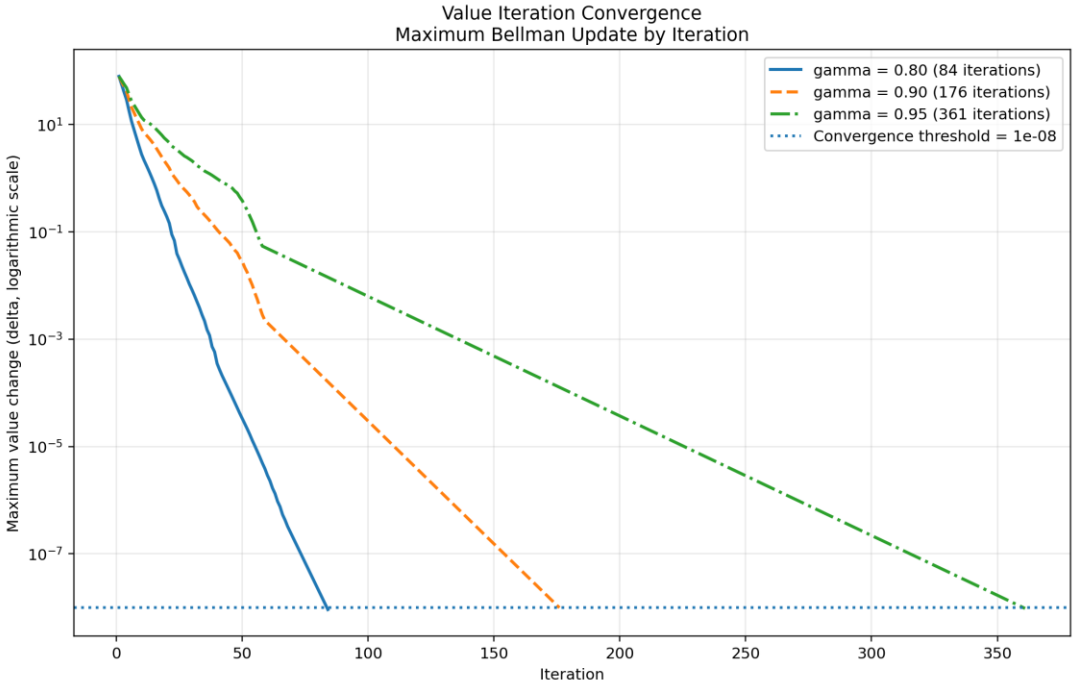
در این Heatmap ، رنگ‌های روشن‌تر نشان‌دهنده مقدار حالت بالاتر و بازده موردانتظار بیشتر هستند. بیشترین مقادیر در ناحیه پایین و نزدیک هدف مشاهده می‌شوند و با دورشدن از هدف، به‌ویژه در بخش بالایی نقشه و اطراف نقطه شروع، مقدار حالت‌ها کاهش می‌یابد. این گرادیان نشان می‌دهد Value Iteration اثر پاداش نهایی هدف را از طریق به‌روزرسانی‌های بلمن به حالت‌های قبلی منتشر کرده است. وجود دیوارها، خانه‌های جریمه، در قفل و دروازه دوره‌ای باعث می‌شود افزایش مقدار حالت کاملاً یکنواخت نباشد و بعضی خانه‌های نزدیک از نظر فاصله هندسی، به دلیل ریسک یا محدودیت عبور، ارزش کمتری داشته باشند. از آنجا که تصویر بیشینه مقدار را روی وضعیت‌های کلید و فاز دروازه نمایش می‌دهد، تفاوت داخلی این وضعیت‌ها در این Heatmap دیده نمی‌شود و برای تحلیل کامل باید سیاست یا مقدار هر زیرحالت نیز جداگانه بررسی شود.



شکل ۲ - سیاست حریصانه **Value Iteration** برای **has_key=0** و**gate_phase=0**

فلش‌های این شکل عمل حریصانه سیاست مرجع Value Iteration را برای حالت has_key=0 و gate_phase=0 نشان می‌دهند. در این وضعیت، عامل هنوز کلید را دریافت نکرده است؛ بنابراین سیاست ابتدا حرکت به سمت کلید و سپس دسترسی به مسیر منتهی به در و هدف را ترجیح می‌دهد. جهت فلش‌ها صرفاً بر اساس کوتاه‌ترین فاصله هندسی تعیین نشده است، بلکه احتمال لغزش عمل، دیوارها، خانه‌های جریمه، در قفل و وضعیت دروازه دوره‌ای نیز در محاسبه ارزش موردانتظار اعمال وارد شده‌اند. در بعضی خانه‌ها ممکن است چند عمل ارزش‌های بسیار نزدیک داشته باشند؛ در نتیجه تغییر کوچک در γ یا نحوه شکستن تساوی می‌تواند جهت فلش را تغییر دهد، بدون آنکه کیفیت کلی مسیر به شکل محسوسی کاهش یابد. این تصویر تنها یکی از چهار ترکیب ممکن وضعیت کلید و فاز دروازه را نمایش می‌دهد و سیاست کامل در فایل مدل ذخیره شده است.

با افزایش gamma ، عامل آینده دورتر را با وزن بیشتری در نظر می‌گیرد و تعداد تکرارهای لازم برای همگرایی افزایش می‌یابد. مدل gamma=0.95 به عنوان سیاست مرجع استفاده شده است، زیرا افق بلندتری دارد و برای مسئله‌ای که نیازمند جمع‌آوری کلید و عبور از چند ناحیه است مناسب‌تر است.



شکل ۳ - مقایسه روند همگرایی Value Iteration برای γ های مختلف

شکل بالا روند همگرایی الگوریتم Value Iteration را برای سه مقدار مختلف ضریب تنزیل نشان می‌دهد. همان‌طور که مشاهده می‌شود، با افزایش مقدار γ تعداد تکرارهای لازم برای همگرایی نیز بیشتر شده است. برای γ=0.80 الگوریتم در ۸۴ تکرار همگرا شده، در حالی که برای γ=0.90 این مقدار به ۱۷۶ تکرار و برای γ=0.95 به ۳۶۱ تکرار رسیده است. این رفتار نشان می‌دهد که مقادیر بزرگ‌تر γ باعث می‌شوند عامل اهمیت بیشتری به پاداش‌های آینده بدهد و در نتیجه انتشار ارزش‌ها در فضای حالت با دقت بیشتری انجام شود، اما سرعت همگرایی کاهش یابد. با این حال، در هر سه حالت مقدار نهایی delta تقریباً در مرتبه ۱۰^-۸ قرار دارد که نشان‌دهنده همگرایی موفق و پایدار الگوریتم است.

ارزیابی نهایی سیاست Value Iteration

به‌منظور ارزیابی عملی سیاست نهایی، الگوریتم Value Iteration با ضریب تنزیل γ=0.95 و آستانه همگرایی $\theta=10^{-8}$ اجرا شد. پس از همگرایی کامل الگوریتم در ۳۶۱ تکرار، سیاست حریصانه حاصل بدون هیچ‌گونه آموزش مجدد یا تغییر تابع ارزش، در ۱۰۰ اپیزود مستقل ارزیابی شد. این ارزیابی روی نقشه اصلی، با پاداش Sparse و Seed های ۱۰۰۰۸ تا ۱۰۱۰۷ انجام گرفت تا اثر تصادفی‌بودن انتقال‌های محیط نیز در نتایج لحاظ شود.

سیاست Value Iteration در تمام ۱۰۰ اپیزود ارزیابی توانست عامل را با موفقیت به هدف برساند و نرخ موفقیت ۱۰۰ درصد را ثبت کند. میانگین پاداش تجمعی در این ارزیابی برابر ۶۵.۷۴ و میانگین تعداد گام لازم برای رسیدن به هدف برابر ۴۷.۶۰ بود. این نتایج نشان می‌دهند که همگرایی عددی الگوریتم صرفاً به تولید مقادیر پایدار محدود نشده است، بلکه سیاست استخراج‌شده در اجرای تصادفی محیط نیز عملکرد پایدار و قابل‌اعتمادی دارد. میانگین نسبتاً پایین تعداد گام‌ها نیز نشان می‌دهد سیاست مرجع توانسته است مسیرهای کارآمدی برای جمع‌آوری کلید، عبور از در و رسیدن به هدف انتخاب کند.

```
Running Value Iteration evaluation
gamma = 0.95
evaluation episodes = 100

Value Iteration evaluation finished.
Converged: True
Iterations: 361
Success rate: 100.00%
Average reward: 65.74
Average steps: 47.60

Episode CSV: C:\Users\pc\Documents\RL_FinalProject_40403384\results\raw_data\value_iteration_evaluation_episodes.csv
Summary CSV: C:\Users\pc\Documents\RL_FinalProject_40403384\results\raw_data\value_iteration_evaluation_summary.csv
JSON: C:\Users\pc\Documents\RL_FinalProject_40403384\results\raw_data\value_iteration_evaluation.json
PS C:\Users\pc\Documents\RL_FinalProject_40403384>
```

الگوریتم	γ	نوع پاداش	تعداد اپیزود	نرخ موفقیت	میانگین پاداش	میانگین گام
Value Iteration	0.95	Sparse	100	100٪	65.74	47.60

نرخ موفقیت ۱۰۰ درصد نشان می‌دهد سیاست Value Iteration در تمام نمونه‌های تصادفی ارزیابی قادر به تکمیل صحیح مأموریت بوده است. تفاوت تعداد گام میان اپیزودها می‌تواند ناشی از تابع انتقال تصادفی محیط باشد؛ زیرا عمل انتخاب‌شده همیشه با قطعیت اجرا نمی‌شود. با وجود این تصادفی‌بودن، میانگین ۴۷.۶۰ گام بیانگر آن است که سیاست نهایی در اکثر اجراها مسیر مناسبی را حفظ کرده است. این ارزیابی همچنین امکان مقایسه کیفیت مسیر Value Iteration با الگوریتم‌های model-free را فراهم می‌کند، مشروط بر آنکه نتایج هر سه روش با نوع پاداش یکسان مقایسه شوند.

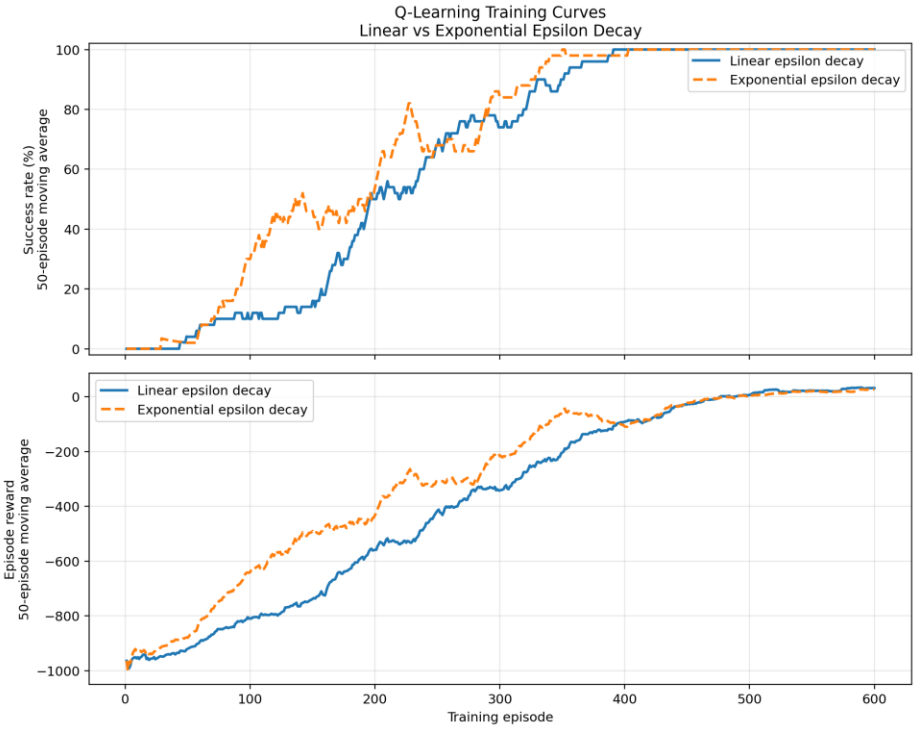
۵. نتایج Q-Learning

نتایج این بخش مربوط به اجرای اصلی Q-Learning با پاداش Shaped ، ضریب کاهش γ=0.95 ، نرخ یادگیری α=0.10 ، Seed پایه ۸ و ۶۰۰ اپیزود آموزش است. مقدار epsilon از ۱.۰۰ آغاز شد و تا حداقل ۰.۰۵ کاهش یافت. پس از پایان آموزش، سیاست حریصانه بدون اکتشاف در ۱۰۰ اپیزود مستقل ارزیابی شد؛ بنابراین نرخ موفقیت، پاداش و تعداد گام‌های جدول مربوط به مرحله ارزیابی نهایی هستند و با مقادیر ثبت‌شده هنگام آموزش تفاوت دارند.

برنامه epsilon	زمان آموزش	موفقیت 100 اپیزود آخر	موفقیت ارزیابی	پاداش ارزیابی	گام ارزیابی
linear	0.89	100.00%	100.00%	9.70	78.67
exponential	0.83	100.00%	100.00%	-32.54	83.65

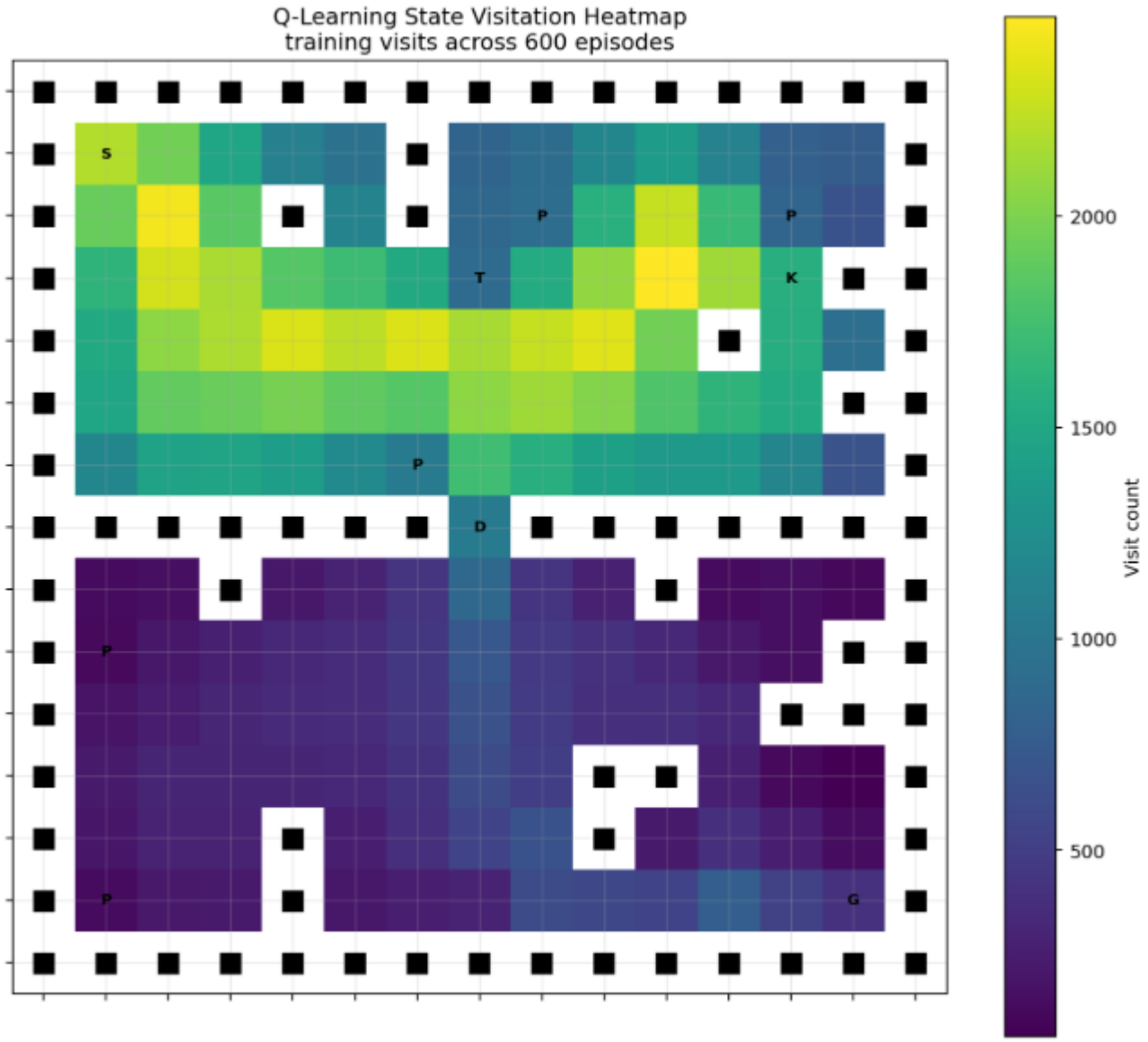
5.0.1 جمع‌بندی نتایج آموزش Q-Learning

در مرحله آموزش، هر دو برنامه کاهش خطی و نمایی ε طی ۶۰۰ اپیزود اجرا شدند. نرخ موفقیت در ۱۰۰ اپیزود پایانی برای هر دو برنامه به ۱۰۰ درصد رسید که نشان می‌دهد هر دو تنظیم در پایان آموزش توانسته‌اند سیاست موفقی یاد بگیرند. بااین‌حال، کیفیت روند آموزش یکسان نبود؛ میانگین متحرک پاداش در ۵۰ اپیزود پایانی برای برنامه خطی برابر 31.82 و برای برنامه نمایی برابر 26.94 بود. بنابراین برنامه خطی در بخش پایانی آموزش، ضمن حفظ نرخ موفقیت کامل، بازده تجمعی بهتری نشان داد. این مقادیر مربوط به مرحله Training هستند و باید از نتایج ارزیابی نهایی بدون اکتشاف تفکیک شوند؛ در مرحله Evaluation ، سیاست حریصانه حاصل از آموزش در ۱۰۰ اپیزود مستقل بررسی شده است.



شکل ۴ ـ روند آموزش Q-Learning با برنامه‌های کاهش خطی و نمایی epsilon بر اساس میانگین متحرک ۵۰ اپیزود

نمودار بالا نرخ موفقیت و پاداش اپیزودی Q-Learning را با میانگین متحرک ۵۰ اپیزود برای دو برنامه کاهش epsilon نشان می‌دهد. هر دو برنامه در پایان آموزش به نرخ موفقیت متحرک ۱۰۰ درصد رسیدند؛ بنابراین هر دو توانستند سیاستی موفق بیاموزند. بااین‌حال، میانگین متحرک پاداش نهایی در کاهش خطی برابر ۳۱.۸۲ و در کاهش نمایی برابر ۲۶.۹۴ بود. این اختلاف نشان می‌دهد برنامه خطی در بخش پایانی آموزش، علاوه بر حفظ موفقیت کامل، بازده تجمعی بهتری ایجاد کرده است و انتخاب آن به‌عنوان تنظیم اصلی Q-Learning را تأیید می‌کند. این مقادیر مربوط به روند آموزش هستند و نباید با پاداش ارزیابی بدون اکتشاف اشتباه گرفته شوند.



شکل 5 نقشه بازدید حالت‌ها توسط **Q-Learning** در طول ۶۰۰ اپیزود آموزش

این نقشه تعداد بازدید عامل Q-Learning از موقعیت‌های مختلف محیط را در طول ۶۰۰ اپیزود آموزش نشان می‌دهد. تعداد بازدید برای حالت‌های دارای وضعیت‌های مختلف has_key وgate_phase در هر مختصات مکانی تجمیع شده است. نواحی با شدت رنگ بیشتر، بخش‌هایی از محیط هستند که عامل در فرایند اکتشاف و بهره‌برداری بیشتر تجربه کرده است. بنابراین این شکل برخلاف ارزیابی حریصانه نهایی، الگوی واقعی بازدید عامل در فرایند یادگیری را نمایش می‌دهد.

در اجرای فعلی، برنامه کاهش خطی epsilon نسبت به برنامه نمایی تعادل بهتری میان اکتشاف اولیه و بهره‌برداری نهایی ایجاد کرد. نقشه بازدید نشان می‌دهد تمرکز عامل روی مسیرهای منتهی به کلید، در و هدف است و خانه‌های خارج از مسیر اصلی کمتر بازدید می‌شوند.

۵،۱. بازسازی یک به‌روزرسانی واقعی Q

ویژگی	مقدار
state	1, 1, 0, 0
action	3
action_name	RIGHT
reward	-0.9600000000000002
next_state	1, 2, 0, 1
terminated	False
truncated	False
old_q	0.0
next_max_q	0.0
target	-0.9600000000000002
td_error	-0.9600000000000002
alpha	0.1
gamma	0.95
new_q	-0.09600000000000003

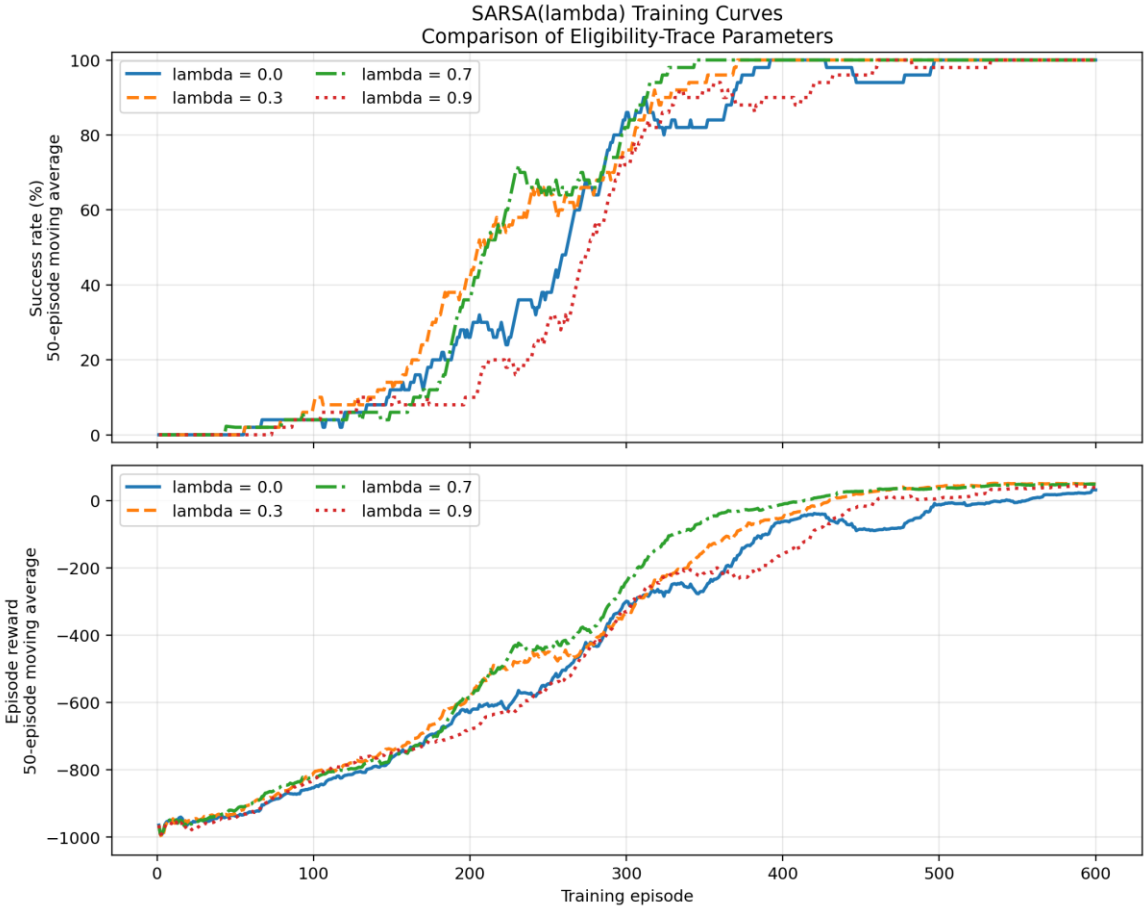
مقادیر جدول بالا مستقیماً از فایل اجرای الگوریتم خوانده شده‌اند. با جایگذاری پاداش، gamma، مقدار بیشینه حالت بعدی و Q قبلی در فرمول، target و Q جدید به دست می‌آیند. این نمونه نشان می‌دهد گزارش عددی به اجرای واقعی متصل است

۶. نتایج SARSA(λ)

نتایج این بخش مربوط به اجرای اصلی SARSA(λ) با پاداش Shaped ، ضریب کاهش γ=0.95 ، نرخ یادگیری α=0.10 ، Seed پایه ۸ و ۶۰۰ اپیزود آموزش است. سیاست رفتار ε-greedy بوده و مقدار epsilon از ۱.۰۰ تا حداقل ۰.۰۵ کاهش یافته است. برای بررسی اثر ردپای شایستگی، مقادیر λ برابر ۰.۰، ۰.۳، ۰.۷ و ۰.۹ با سایر تنظیمات یکسان آزمایش شدند. پس از پایان آموزش، سیاست حریصانه هر مدل بدون اکتشاف در ۱۰۰ اپیزود مستقل ارزیابی شد. بنابراین ستون‌های موفقیت ارزیابی، پاداش و تعداد گام، عملکرد سیاست نهایی را نشان می‌دهند و مستقیماً معیارهای دوره آموزش نیستند.

lambda	زمان آموزش	موفقیت 100 اپیزود آخر	موفقیت ارزیابی	پاداش	گام
0.0	0.90	100.00%	99.00%	-6.66	92.61
0.3	1.57	100.00%	100.00%	56.02	49.84
0.7	2.46	100.00%	100.00%	53.45	51.68
0.9	4.64	100.00%	98.00%	35.58	64.29

بر اساس نتایج فعلی، lambda=0.3 بهترین تعادل کلی میان نرخ موفقیت، پاداش، طول مسیر و زمان آموزش را ایجاد می‌کند. مقادیر بزرگ‌تر lambda اطلاعات بیشتری را به گذشته منتقل می‌کنند، اما هزینه محاسباتی و حساسیت به مسیرهای تصادفی را نیز افزایش می‌دهند.



شکل ۶ - منحنی‌های یادگیری SARSA(λ) برای مقادیر مختلف λ

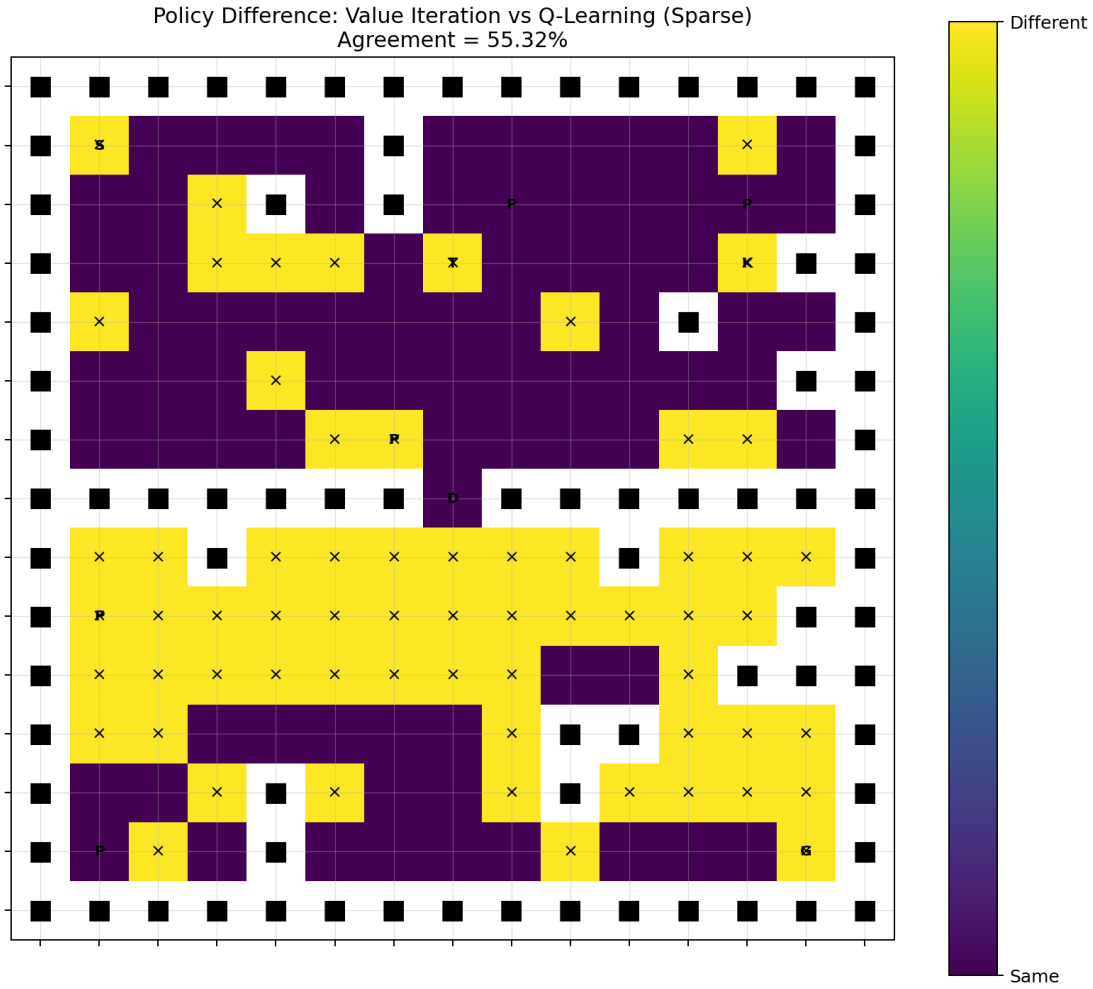
نمودار منحنی‌های یادگیری SARSA(λ) نشان می‌دهد که هر چهار مقدار λ در نهایت به نرخ موفقیت ۱۰۰ درصد می‌رسند، اما سرعت یادگیری و کیفیت پاداش نهایی آن‌ها متفاوت است. بر اساس نتایج، λ=0.3 بهترین عملکرد را از نظر میانگین پاداش نهایی نشان می‌دهد و به تعادل مناسب‌تری بین انتشار اعتبار در گام‌های متوالی و پایداری یادگیری می‌رسد. در مقابل، λ=0.0 رفتار نزدیک‌تر به TD تک‌مرحله‌ای دارد و اگرچه پایدار است، پاداش نهایی کمتری نسبت به λ=0.3 به دست می‌آورد. همچنین مقادیر بزرگ‌تر مانند λ=0.9 باعث انتقال گسترده‌تر اعتبار می‌شوند، اما در این مسئله به بهترین عملکرد منجر نشده‌اند.

۶،۱. در یک اپیزود نمونه E و delta تغییرات

گام	حالت	عمل	delta	تعداد ردپا	پس از کاهش E
1	[1, 1, 0, 0]	3	-0.9600	1	RIGHT:0.6650
2	[1, 2, 0, 1]	0	-0.9700	2	UP:0.6650, RIGHT:0.4422
3	[1, 3, 0, 0]	1	-0.9800	3	DOWN:0.6650, UP:0.4422, RIGHT:0.2941
4	[2, 3, 0, 1]	1	-0.9900	4	DOWN:0.6650, DOWN:0.4422, UP:0.2941, RIGHT:0.1956
5	[3, 3, 0, 0]	0	-1.3800	5	UP:0.6650, DOWN:0.4422, DOWN:0.2941, UP:0.1956, RIGHT:0.1300
6	[2, 3, 0, 1]	3	-5.1800	6	RIGHT:0.6650, UP:0.4422, DOWN:0.2941, DOWN:0.1956, UP:0.1300, RIGHT:0.0865
7	[2, 3, 0, 0]	1	-1.3700	7	DOWN:0.6650, RIGHT:0.4422, UP:0.2941, DOWN:0.1956, DOWN:0.1300, UP:0.0865, RIGHT:0.0575
8	[2, 2, 0, 1]	3	-1.1102	8	RIGHT:0.6650, DOWN:0.4422, RIGHT:0.2941, UP:0.1956, DOWN:0.1300, DOWN:0.0865, UP:0.0575, RIGHT:0.0382

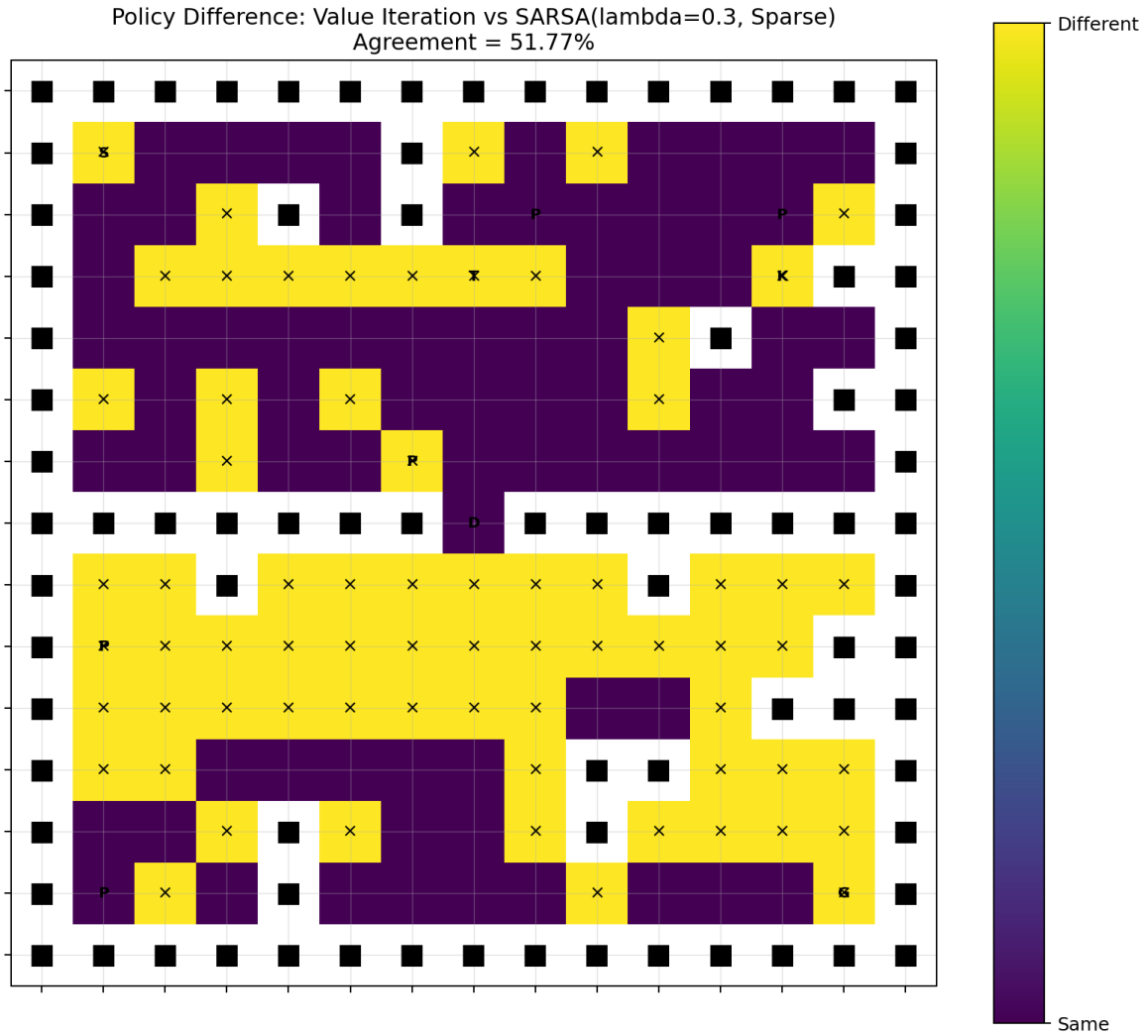
علامت و اندازه delta جهت اصلاح Q را مشخص می‌کند. ردپاهای E پس از هر گام با gamma*lambda کاهش می‌یابند و باعث می‌شوند خطای جدید روی چند زوج حالت-عمل قبلی نیز اثر بگذارد. در replacing trace، ردپای زوج جاری به مقدار یک جایگزین می‌شود.

۷. مقایسه سیاست‌ها و تحلیل اختلاف



شکل ۷ - نقشه اختلاف سیاست Value Iteration و Q-Learning با پاداش Sparse

در مقایسه روی پاداش یکسان Sparse و Seed پایه ۸، درصد توافق سیاست Q-Learning با سیاست مرجع Value Iteration برابر ۵۵.۳۲ درصد است. خانه‌های تیره نشان‌دهنده عمل حریصانه یکسان و خانه‌های روشن نشان‌دهنده اختلاف سیاست هستند. این اختلاف می‌تواند از نمونه‌برداری محدود، انتقال تصادفی و نزدیک‌بودن مقدار چند عمل در بعضی حالت‌ها ناشی شود.



شکل ۸ - نقشه اختلاف سیاست **Value Iteration** و **SARSA(lambda=0.3)** با پاداش**Sparse**

درصد توافق سیاست SARSA(lambda=0.3) با Value Iteration در همان نقشه، پاداش Sparse و Seed پایه ۸ برابر ۵۱.۷۷ درصد است. این مقدار حدود ۳.۵۵ واحد درصد کمتر از Q-Learning است. با وجود موفقیت ارزیابی بالای SARSA ، توافق کمتر سیاست‌ها لزوماً به معنای مسیر ضعیف‌تر نیست؛ زیرا چند عمل متفاوت ممکن است در محیط تصادفی به مسیرهایی با کیفیت نزدیک منتهی شوند .

۷,۱. مقایسه مستقیم سه الگوریتم با پاداش Sparse

روش	زمان اجرا	نمونه یا محاسبه	پایداری	حافظه	کیفیت مسیر	حساسیت پارامتری
Value Iteration (gamma=0.95)	5.20 ثانیه	پیمایش کامل 361 فضای حالت	قطعی در مدل و تنظیمات ثابت	Peak Memory = 0.1348 MB (tracemalloc)	موفقیت ۱۰۰٪؛ میانگین پاداش ٦٥,٧٤؛ میانگین مسیر ٤٧,٦٠ گام	gamma و آستانه همگرایی
Q-Learning / Sparse	0.83 ± 0.02 ثانیه	3 × اجرای مستقل اپیزود 600	موفقیت 98.67 ±1.53 درصد؛ گام 65.46 ± 2.98	Peak Memory = 0.7397 MB (tracemalloc)	میانگین مسیر 65.46 گام؛ موفقیت 98.67 درصد	alpha، gamma، epsilon و برنامه کاهش
SARSA(lambda=0.3) / Sparse	1.56 ± 0.03 ثانیه	3 × اجرای مستقل اپیزود 600	موفقیت 100 ±0 درصد؛ گام 54.17 ± 1.83	Peak Memory = 0.7132 MB (tracemalloc)	میانگین مسیر 54.17 گام؛ موفقیت 100 درصد	alpha، gamma، epsilon، lambda و نوع ردپا

زمان اجرای این روش‌ها مستقیماً هم‌ارز نیست؛ زیرا Value Iteration از پیمایش کامل مدل انتقال استفاده می‌کند، در حالی که Q-Learning و SARSA از اپیزودهای نمونه‌برداری‌شده می‌آموزند. با این حال، جدول تصویری فشرده از هزینه محاسباتی، حافظه، پایداری و کیفیت سیاست نهایی ارائه می‌دهد.

مقایسه حافظه: علاوه بر مقایسه ساختارهای ذخیره‌سازی، مصرف حافظه در اجرای واقعی سه الگوریتم نیز با ابزار tracemalloc پایتون و روی نقشه و تابع پاداش یکسان اندازه‌گیری شد. مقدار بیشینه‌ی تخصیص حافظه‌ی قابل‌ردیابی در اجرای Value Iteration برابر با 0.1348 MB، در Q-Learning برابر با 0.7397MB و در SARSA(λ=0.3) برابر با 0.7132 MB بود. از نظر ساختار جدولی، Value Iteration برای ۵۶۴ حالت مقادیر و سیاست را نگهداری می‌کند، در حالی که Q-Learning و SARSA برای ۵۶۴ حالت و ۴ عمل دارای حداکثر مقدار Q هستند. در SARSA(λ)، ردپای شایستگی نیز در طول هر اپیزود نگهداری می‌شود، اما در این پیاده‌سازی این ساختار به‌صورت sparse و موقت است و همه‌ی ۲۲۵۶ مؤلفه به‌طور هم‌زمان ذخیره نمی‌شوند؛ به همین دلیل مصرف اوج اندازه‌گیری‌شده‌ی SARSA در این اجرا اندکی کمتر از Q-Learning بود. این مقادیر مصرف کل RAM سیستم‌عامل نیستند، بلکه بیشینه‌ی تخصیص حافظه‌ی مدیریت‌شده توسط پایتون هستند که با tracemalloc ثبت شده‌اند. داده‌ی خام این آزمایش نیز در فایل results/raw_data/algorithm_memory_comparison.csv ذخیره شده است.

۷,۲. سه حالت نمونه با سیاست متفاوت

ساختار محلی	عمل Q-Learning	عمل Value Iteration	حالت
.=بالا#=، پایین.=، چپ#=، راست	راست (RIGHT)	پایین (DOWN)	(1, 1, 0, 0)
.=بالا#=، پایین.=، چپ.=، راست	چپ (LEFT)	پایین (DOWN)	(1, 3, 0, 0)
.=بالا#=، پایین#=، چپ.=، راست	چپ (LEFT)	راست (RIGHT)	(1, 4, 0, 0)

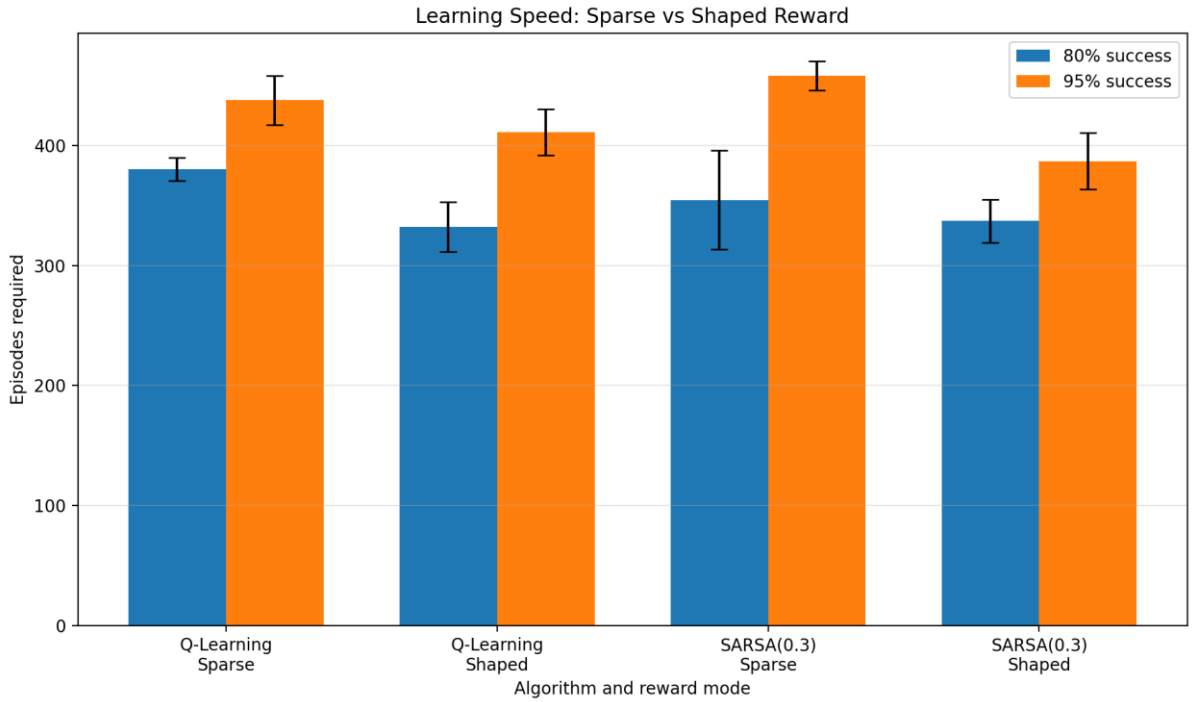
در این حالت‌ها وجود دیوار، گلوگاه، خانه جریمه یا چند مسیر هم‌ارزش می‌تواند باعث اختلاف شود Value Iteration. از مدل کامل استفاده می‌کند، ولی Q-Learning تنها بر نمونه‌های تجربه‌شده تکیه دارد؛ بنابراین در حالت‌های کم بازدید، اختلاف طبیعی است.

۷,۳. مقایسه پاداش Sparse و Shaped

برای بررسی اثر شکل‌دهی پاداش و پایداری نتایج، Q-Learning و SARSA(λ=0.3) در دو حالت Sparse و Shaped با Seed های مستقل ۸، ۱۸ و ۲۸ و در ۶۰۰ اپیزود اجرا شدند. تمام مقادیر جدول به‌صورت میانگین ± یک انحراف معیار گزارش شده‌اند.

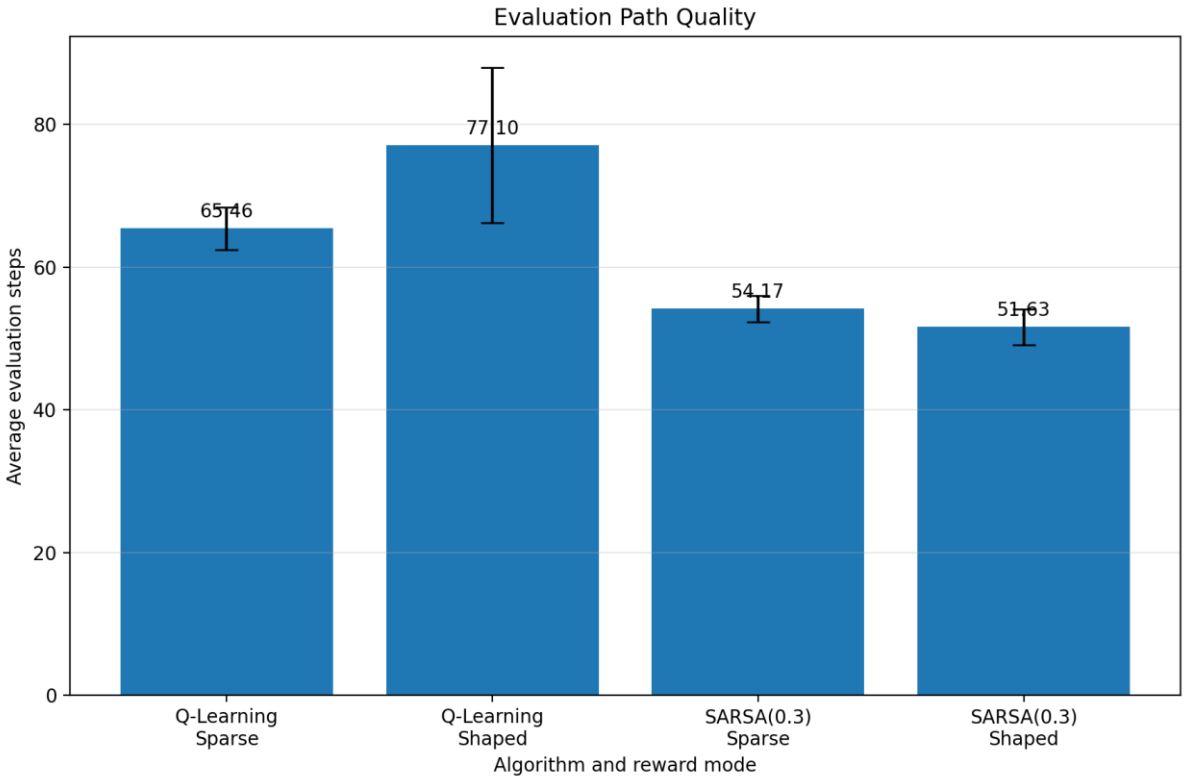
روش	موفقیت ارزیابی	میانگین گام	%اپیزود تا 80	%اپیزود تا 95	زمان آموزش
Q-Learning / Sparse	98.67 ± 1.53%	65.46 ± 2.98	380.33 ± 9.61	437.67 ± 20.55	0.83 ± 0.02
Q-Learning / Shaped	100.00 ± 0.00%	77.10 ± 10.87	332.33 ± 20.79	411.00 ± 19.08	0.86 ± 0.06
SARSA(λ=0.3) / Sparse	100.00 ± 0.00%	54.17 ± 1.83	354.67 ± 41.00	458.00 ± 12.12	1.56 ± 0.03
SARSA(λ=0.3)	100.00 ± 0.00%	51.63 ± 2.49	337.00 ± 17.78	387.00 ± 23.39	1.52 ± 0.00

/ Shaped					



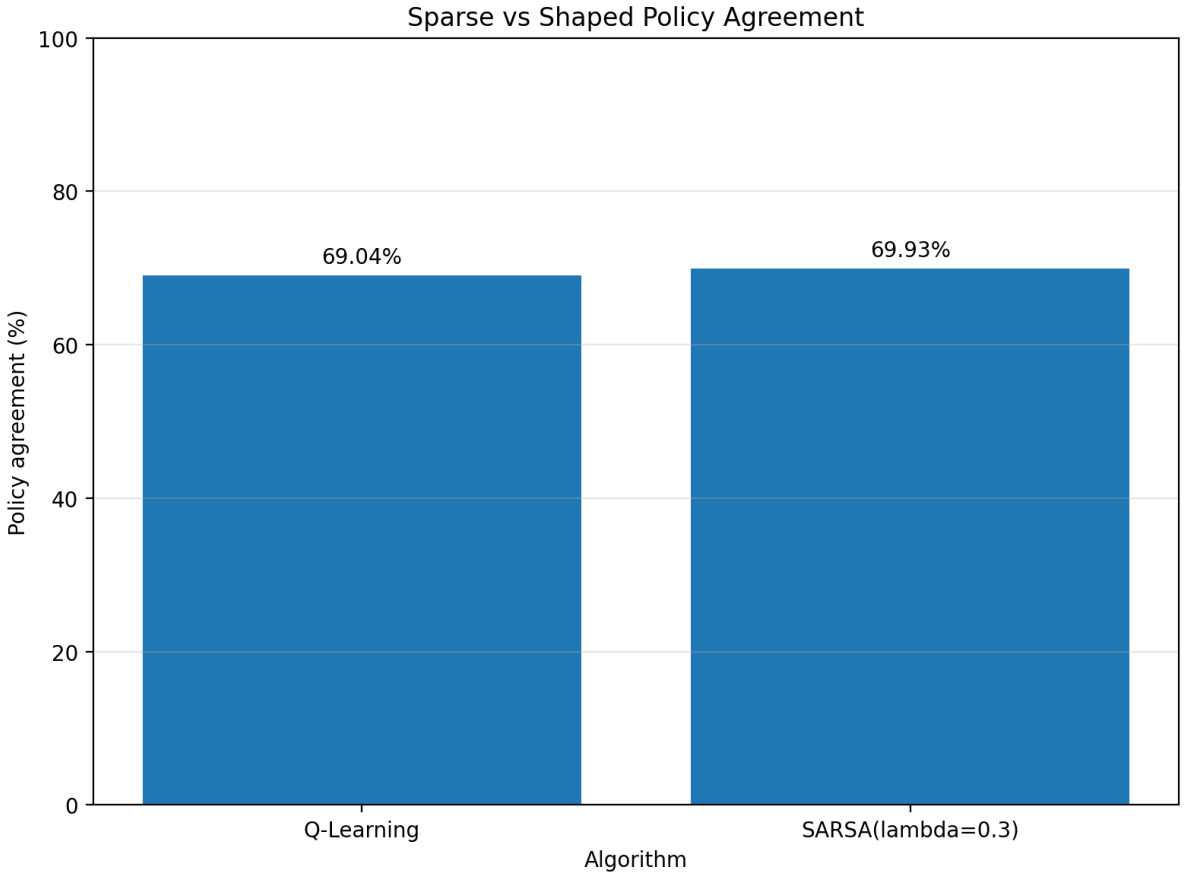
نمودار تکمیلی الف - سرعت یادگیری در پاداش **Sparse** و**Shaped** ؛ خطوط خطا یک انحراف معیار سه اجرای مستقل هستند.

این نمودار نشان می‌دهد پاداش Shaped در هر دو الگوریتم باعث رسیدن سریع‌تر عامل به آستانه‌های موفقیت شده است. در Q-Learning، میانگین اپیزود لازم برای رسیدن به موفقیت ۸۰ درصد از ۳۸۰.۳۳ در حالت Sparse به ۳۳۲.۳۳ کاهش یافت و آستانه ۹۵ درصد نیز به‌جای ۴۳۷.۶۷ در اپیزود ۴۱۱ حاصل شد. در SARSA(λ=0.3) ، اثر شکل‌دهی پاداش در آستانه ۹۵ درصد واضح‌تر بود و تعداد اپیزودها از ۴۵۸ به ۳۸۷ کاهش یافت. پراکندگی بیشتر نتایج SARSA در حالتSparse ، به‌ویژه انحراف معیار ۴۱ اپیزود برای آستانه ۸۰ درصد، نشان می‌دهد سرعت یادگیری بدون بازخورد میانی به مسیرهای تصادفی و Seed حساس‌تر بوده است. بااین‌حال، سریع‌ترشدن یادگیری به‌تنهایی اثبات نمی‌کند که سیاست نهایی کوتاه‌تر یا به سیاست بهینه نظری نزدیک‌تر است؛ کیفیت مسیر و توافق سیاست باید جداگانه بررسی شوند.



نمودار تکمیلی ب - کیفیت مسیر در مرحله ارزیابی بر اساس میانگین تعداد گام‌ها؛ مقادیر کمتر نشان‌دهنده مسیرهای کوتاه‌تر هستند.

این نمودار کیفیت مسیر سیاست نهایی را با میانگین تعداد گام لازم برای تکمیل موفق مأموریت مقایسه می‌کند؛ بنابراین مقدار کمتر نشان‌دهنده مسیر کوتاه‌تر و کارآمدتر است. در Q-Learning ، میانگین طول مسیر در حالت Sparse برابر 65.46 ± 2.98 گام بود، در حالی که در حالت Shaped به 77.10 ± 10.87 گام افزایش یافت. این نتیجه نشان می‌دهد پاداش شکل‌داده‌شده اگرچه سرعت یادگیری را افزایش داده است، در بودجه ۶۰۰ اپیزودی الزاماً به مسیر نهایی کوتاه‌تر در Q-Learning منجر نشده و پراکندگی بیشتری نیز میان Seed ها ایجاد کرده است. در مقابل، برای SARSA(λ=0.3) ، میانگین مسیر از 54.17 ± 1.83 گام در حالت Sparse به 51.63 ± 2.49 گام در حالت Shaped کاهش یافته است. بنابراین اثر شکل‌دهی پاداش به الگوریتم وابسته بوده و باید سرعت یادگیری و کیفیت نهایی مسیر به‌صورت دو معیار مستقل بررسی شوند. همچنین به دلیل تصادفی‌بودن انتقال‌ها، تعداد گام مشاهده‌شده فقط تابع طول هندسی مسیر نیست و لغزش عمل یا انتظار برای بازشدن دروازه نیز بر آن اثر می‌گذارد.



نمودار تکمیلی پ - درصد توافق سیاست‌های **Sparse** و **Shaped** برای مدل‌های **Seed** پایه ۸.

این نمودار میزان یکسان‌بودن عمل حریصانه سیاست‌های آموزش‌دیده با پاداش Sparse و Shaped را برای مدل‌های Seed پایه ۸ نشان می‌دهد. درصد توافق برای Q-Learning برابر ۶۹.۰۴ درصد و برای SARSA(λ=0.3) برابر ۶۹.۹۳ درصد است؛ بنابراین در حدود ۳۰ درصد از حالت‌های مقایسه‌شده، نوع تابع پاداش موجب انتخاب عمل حریصانه متفاوت شده است. این اختلاف می‌تواند ناشی از بازخورد میانی پاداشShaped ، بودجه محدود ۶۰۰ اپیزودی، بازدید نامتوازن حالت‌ها و نزدیک‌بودن مقدار چند عمل باشد. توافق کمتر از ۱۰۰ درصد الزاماً نشان‌دهنده ضعیف‌بودن یکی از سیاست‌ها نیست، زیرا چند مسیر متفاوت ممکن است نرخ موفقیت و طول مسیر مشابهی ایجاد کنند. همچنین این مقایسه تنها سیاست‌های یادگرفته‌شده در یک Seed مشخص را بررسی می‌کند و اثبات نمی‌کند که شکل‌دهی پاداش، سیاست بهینه نظری را تغییر داده است.

در هر دو الگوریتم، استفاده از پاداش **Shaped** باعث شد عامل سریع‌تر به نرخ موفقیت ۸۰ و ۹۵ درصد برسد. در **Q-Learning**، مسیرهای ارزیابی در حالت Shaped از نظر میانگین تعداد گام‌ها طولانی‌تر و دارای پراکندگی بیشتری بودند؛ در حالی که در **SARSA**، حالت Shaped به مسیرهایی کمی کوتاه‌تر منجر شد. درصد توافق سیاست‌های Sparse و Shaped برای Q-Learning برابر با **۶۹.۰۴٪** و برای SARSA برابر با **۶۹.۹۳٪** بود. این میزان اختلاف نشان می‌دهد که در بودجه محدود **۶۰۰ اپیزود**، شکل‌دهی پاداش علاوه بر افزایش سرعت یادگیری، می‌تواند سیاست حریصانه یادگرفته‌شده را نیز تغییر دهد. با این حال، بر اساس این آزمایش نمی‌توان نتیجه گرفت که سیاست بهینه نظری مسئله الزاماً تغییر کرده است. همچنین، پاداش عددی دو حالت به‌صورت مستقیم مقایسه نشد، زیرا مقیاس و تعریف پاداش در حالت‌های Sparse و Shaped متفاوت است.

۷,٤. بررسی رفتارهای ناخواسته ناشی از Reward Shaping

برای بررسی رفتارهای ناخواسته ناشی از Reward Shaping ، عملکرد عامل‌ها در ۱۰۰ اپیزود پایانی آموزش بررسی شد. در همه اجراهای Sparse و Shaped نرخ موفقیت ۱۰۰٪ و نرخ پایان به علت محدودیت گام (timeout) صفر بود. همچنین هیچ اپیزود موفقی مشاهده نشد که هم مسیر بسیار طولانی داشته باشد و هم پاداش غیرعادی بالایی دریافت کند (high_reward_long_count = 0). بنابراین شواهدی از reward farming ، طولانی‌کردن عمدی مسیر برای کسب پاداش بیشتر یا رفتار حلقه‌ای پایدار در انتهای آموزش مشاهده نشد.

برای بررسی اجتناب بیش‌ازحد از نواحی خطرناک نیز تعداد برخورد با دیوار و مراجعه به خانه‌های جریمه‌دار مقایسه شد. در Q-Learning میانگین تعداد مراجعه به خانه‌های جریمه‌دار در ۱۰۰ اپیزود پایانی برای Shaped حدود ۰.۳۷ و برای Sparse حدود ۰.۴۰ بار در هر اپیزود بود؛ در SARSA(λ=0.3) نیز این مقادیر به‌ترتیب حدود ۰.۲۹ و ۰.۲۶ بود. بنابراین Reward Shaping باعث حذف غیرعادی یا اجتناب افراطی از این نواحی نشده است.

همچنین اثر Reward Shaping بر طول مسیر به الگوریتم وابسته بود؛ در Q-Learning میانگین طول مسیر در حالت Shaped بیشتر از Sparse شد، در حالی که در SARSA(λ=0.3) اندکی کاهش یافت. با وجود این، نبود timeout و نبود اپیزودهای موفق با مسیر بسیار طولانی و پاداش غیرعادی بالا نشان می‌دهد شواهدی از حلقه‌های پایدار یا reward farming در انتهای آموزش مشاهده نشده است.

۷,٥. مقایسه رفتار Off-policy و On-policy در نزدیکی خطر

Q-Learning پدر محاسبه هدف به‌روزرسانی از بیشینه مقدار Q در حالت بعدی استفاده می‌کند و عمل اکتشافی بعدی را مستقیماً در هدف لحاظ نمی‌کند؛ بنابراین یک روش off-policy است. در مقابل، SARSA از همان عمل ε-greedy که عامل در حالت بعدی انتخاب می‌کند استفاده می‌کند و ریسک ناشی از اکتشاف را نیز در ارزش سیاست وارد می‌سازد. در ۱۰۰ اپیزود آخر اجرای Shaped با Seed پایه ۸، میانگین برخورد با دیوار در Q-Learning خطی برابر ۳.۵۴ و ورود به خانه جریمه برابر ۰.۴۱ بود؛ این مقادیر برای SARSA(λ=0.3) به‌ترتیب ۲.۱۵ و ۰.۱۹ ثبت شدند. میانگین طول اپیزود نیز از ۶۶.۵ گام در Q-Learning به ۵۳.۵ گام در SARSA کاهش یافت. این نتایج با رفتار محافظه‌کارانه‌تر SARSA در نزدیکی خانه‌های خطرناک سازگار است، هرچند نتیجه به نقشه، Seed و پارامترهای همین آزمایش محدود است.

۸. انتقال یادگیری

دو محیط مقصد طراحی شده‌اند. در محیط مشابه، جایگاه شروع، کلید و هدف ثابت مانده و نزدیک‌ترین تعداد صحیح از دیوارهای قابل جابه‌جایی، متناسب با بازه موردنظر، جابه‌جا شده است. در محیط متفاوت، درصد بیشتری از دیوارها تغییر کرده، موقعیت کلید نیز عوض شده و سه خانه جریمه به نقشه افزوده شده‌اند. اعتبار هر دو نقشه با الگوریتم BFS بررسی شده است.

روش‌های انتقال و معیار ارزیابی

در سناریوی Scratch ، جدول Q بدون هیچ دانش اولیه و با مقادیر صفر آموزش داده شد و به‌عنوان خط مبنا در نظر گرفته شد. در انتقال Full ، تمام مقادیر Q قابل‌تطبیق از مدل مبدأ مستقیماً به محیط مقصد منتقل شدند. در انتقال Scaled ، مقادیر منتقل‌شده در ضرایب ۰.۲۵، ۰.۵۰ و ۰.۷۵ ضرب شدند تا شدت اثر دانش قبلی کنترل شود. در روش Selective ، تنها حالت‌هایی منتقل شدند که ساختار همسایگی محلی آن‌ها در نقشه مبدأ و مقصد یکسان بود؛ حالت‌های مجاور دیوارها یا نواحی تغییرکرده حذف شدند تا احتمال انتقال دانش نامتناسب کاهش یابد.

عملکرد هر سناریو در سه مرحله بررسی شد: پیش از آموزش مقصد، طی ۱۰۰ اپیزود ابتدایی و پس از پایان آموزش. معیارهای ارزیابی شامل نرخ موفقیت، میانگین پاداش و میانگین تعداد گام بودند. در این پروژه، انتقال منفی زمانی ثبت شد که یک روش انتقالی نسبت به آموزش از صفر، در موفقیت اولیه یا میانگین پاداش ۱۰۰ اپیزود اول عملکرد ضعیف‌تری ایجاد کند. این تعریف امکان شناسایی حالتی را فراهم می‌کند که دانش قبلی ممکن است در ابتدا مفید به نظر برسد، اما روند یادگیری اولیه را از نظر پاداش مختل کند.

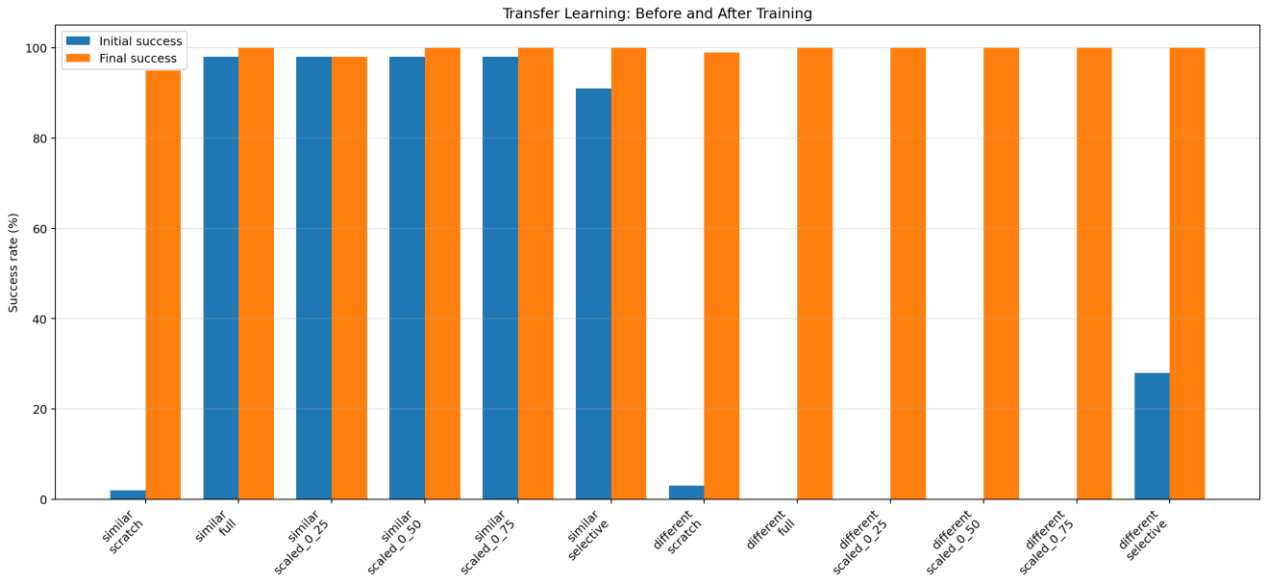
مقصد	سناریو	موفقیت اولیه	اپیزود 100 اول	موفقیت نهایی	پاداش	گام	انتقال منفی
similar	scratch	2.00%	4.00%	95.00%	-35.70	110.43	False
similar	full	98.00%	39.00%	100.00%	58.01	48.33	False
similar	scaled_0_25	98.00%	37.00%	98.00%	31.35	66.92	False
similar	scaled_0_50	98.00%	41.00%	100.00%	56.93	47.91	False
similar	scaled_0_75	98.00%	34.00%	100.00%	58.20	47.78	False
similar	selective	91.00%	26.00%	100.00%	57.04	48.22	False
different	scratch	3.00%	32.00%	99.00%	40.74	58.42	False
different	full	0.00%	52.00%	100.00%	77.47	32.37	True
different	scaled_0_25	0.00%	42.00%	100.00%	78.11	32.33	True
different	scaled_0_50	0.00%	53.00%	100.00%	77.87	32.51	True
different	scaled_0_75	0.00%	56.00%	100.00%	77.77	32.26	True
different	selective	28.00%	23.00%	100.00%	74.85	33.05	True

نتایج نشان می‌دهند اثر انتقال به میزان شباهت محیط مبدأ و مقصد وابسته است. در مقصد مشابه، آموزش از صفر تنها با موفقیت اولیه ۲ درصد آغاز شد، در حالی که انتقال کامل و انتقال‌های مقیاس‌شده موفقیت اولیه ۹۸ درصد ایجاد کردند. انتقال انتخابی نیز با موفقیت اولیه ۹۱ درصد شروع شد. بنابراین دانش مبدأ در محیط مشابه تا حد زیادی معتبر باقی مانده و موجب حذف بخش قابل‌توجهی از مرحله اکتشاف اولیه شده است.

در مقصد متفاوت، رفتار انتقال پیچیده‌تر بود. آموزش از صفر موفقیت اولیه ۳ درصد داشت، اما انتقال کامل و هر سه انتقال مقیاس‌شده با به‌تنهایی برای حذف جهت‌گیری نامناسب سیاست Q موفقیت اولیه صفر درصد آغاز شدند. این نتیجه نشان می‌دهد کاهش اندازه مقادیر منتقل‌شده کافی نبوده است؛ زیرا حتی مقیاس ۰.۲۵ نیز در انتخاب عمل حریصانه اولیه اثر گذاشته است. انتقال انتخابی با حذف حالت‌های دارای ساختار محلی تغییرکرده، موفقیت اولیه را به ۲۸ درصد افزایش داد و از نظر شروع اولیه از همه روش‌ها بهتر بود.

با وجود این شروع مناسب، موفقیت ۱۰۰ اپیزود اول در انتقال انتخابی مقصد متفاوت برابر ۲۳ درصد و در آموزش از صفر برابر ۳۲ درصد بود. علاوه بر آن، میانگین پاداش ۱۰۰ اپیزود اول انتقال انتخابی ۴۰.۹۰ واحد کمتر از آموزش صفر ثبت شد؛ بنابراین طبق معیار تعریف‌شده، این سناریو همچنان نمونه انتقال منفی محسوب می‌شود. در پایان آموزش، تمام روش‌های انتقالی مقصد متفاوت به موفقیت ۱۰۰ درصد رسیدند و میانگین مسیر آن‌ها حدود ۳۲ تا ۳۳ گام بود، در حالی که آموزش از صفر با موفقیت ۹۹ درصد به میانگین ۵۸.۴۲

گام رسید. این نتیجه نشان می‌دهد دانش منتقل‌شده ممکن است در کوتاه‌مدت زیان‌آور باشد، اما پس از اصلاح با تجربه‌های مقصد، همچنان به سیاست نهایی کارآمدتری منتهی شود.



شکل ۹ - مقایسه نرخ موفقیت پیش و پس از آموزش در سناریوهای انتقال یادگیری

در محیط مشابه، روش‌های انتقال کامل و انتقال مقیاس‌شده عملکرد اولیه بسیار بالایی نشان می‌دهند و نسبت به آموزش از صفر، شروع سریع‌تری دارند. در محیط متفاوت، انتقال کامل و برخی ضرایب انتقال مقیاس‌شده ممکن است در ابتدا عملکردی نزدیک به صفر داشته باشند که نمونه‌ای از انتقال منفی محسوب می‌شود. انتقال انتخابی، با حذف حالت‌هایی که همسایگی محلی آن‌ها تغییر کرده است، شروع ایمن‌تری فراهم می‌کند. با ادامه آموزش، همه سناریوهای ثبت‌شده به عملکرد نهایی بالایی می‌رسند؛ با این حال، سرعت یادگیری و کیفیت عملکرد اولیه در آن‌ها متفاوت است.

۸,۱. Q نمونه عددی انتقال منفی و اصلاح

شاخص	مقدار	تفسیر
حالت مقصد	different	نقشه مقصد دارای تغییرات ساختاری بیشتر است.
روش انتقال	selective	فقط حالت‌های دارای همسایگی محلی مشابه منتقل شدند.
حالت‌های منتقل‌شده	180	درصد از حالت‌های ممکن 31.91
حالت‌های حذف‌شده	340	به علت تغییر همسایگی محلی منتقل نشدند.
موفقیت اولیه	28.00%	واحد درصد نسبت به آموزش از صفر +25.00
پاداش 100 اپیزود اول	-822.05	نسبت به آموزش از صفر -40.90
موفقیت نهایی	100.00%	عامل با ادامه آموزش به موفقیت کامل رسید.
پاداش نهایی	74.85	نسبت به آموزش از صفر +34.11
میانگین گام نهایی	33.05	کیفیت مسیر پس از اصلاح سیاست

در سناریوی مقصد متفاوت با استفاده از روش انتقال انتخابی، نرخ موفقیت اولیه ۲۵ واحد درصد بهتر از سناریوی آموزش از صفر بود؛ با این حال، میانگین پاداش تجمعی در ۱۰۰ اپیزود اول، ۴۰.۹۰ واحد کمتر از آموزش از صفر ثبت گردید. به همین دلیل، مطابق با معیارهای تعریف‌شده در کدهای پروژه، این پدیده به عنوان «انتقال منفی (Negative Transfer)» تشخیص داده شد. با ادامه فرآیند آموزش، عامل موفق شد اثر نامناسب دانش اولیه انتقالی را اصلاح کند، به طوری که در نهایت نرخ موفقیت به ۱۰۰ درصد و میانگین پاداش نهایی به ۷۴.۸۵ رسید.

۸،۱،۱. نمونه‌ی state-level از انتقال منفی و اصلاح رفتار

برای نمایش انتقال منفی در سطح یک حالت مشخص، حالت در نقشه مقصد مشابه بررسی شد. در مدل مبدأ، مقادیر Q این حالت برابر با بود و بنابراین عمل حریصانه‌ی منتقل‌شده **UP** انتخاب می‌شد. با این حال، همسایگی محلی این حالت در نقشه مقصد تغییر کرده بود (local_neighborhood_changed = true) و در ساختار جدید، بهترین عمل فوری **RIGHT** بود؛ پاداش موردانتظار RIGHT برابر با و برای UP برابر با بود. بنابراین دانش منتقل‌شده در این حالت مشخص، به انتخاب عملی نامناسب نسبت به ساختار جدید منجر می‌شد و یک نمونه‌ی مستقیم از Negative Transfer است.

پس از ادامه‌ی آموزش روی نقشه مقصد، مقادیر Q همین حالت به تغییر کرد و عمل حریصانه از **UP** به **RIGHT** اصلاح شد (best_action = 3). این تغییر نشان می‌دهد عامل توانسته اثر دانش نامناسب منتقل‌شده را با تجربه‌ی جدید جبران کند و سیاست خود را با ساختار واقعی محیط مقصد سازگار سازد.

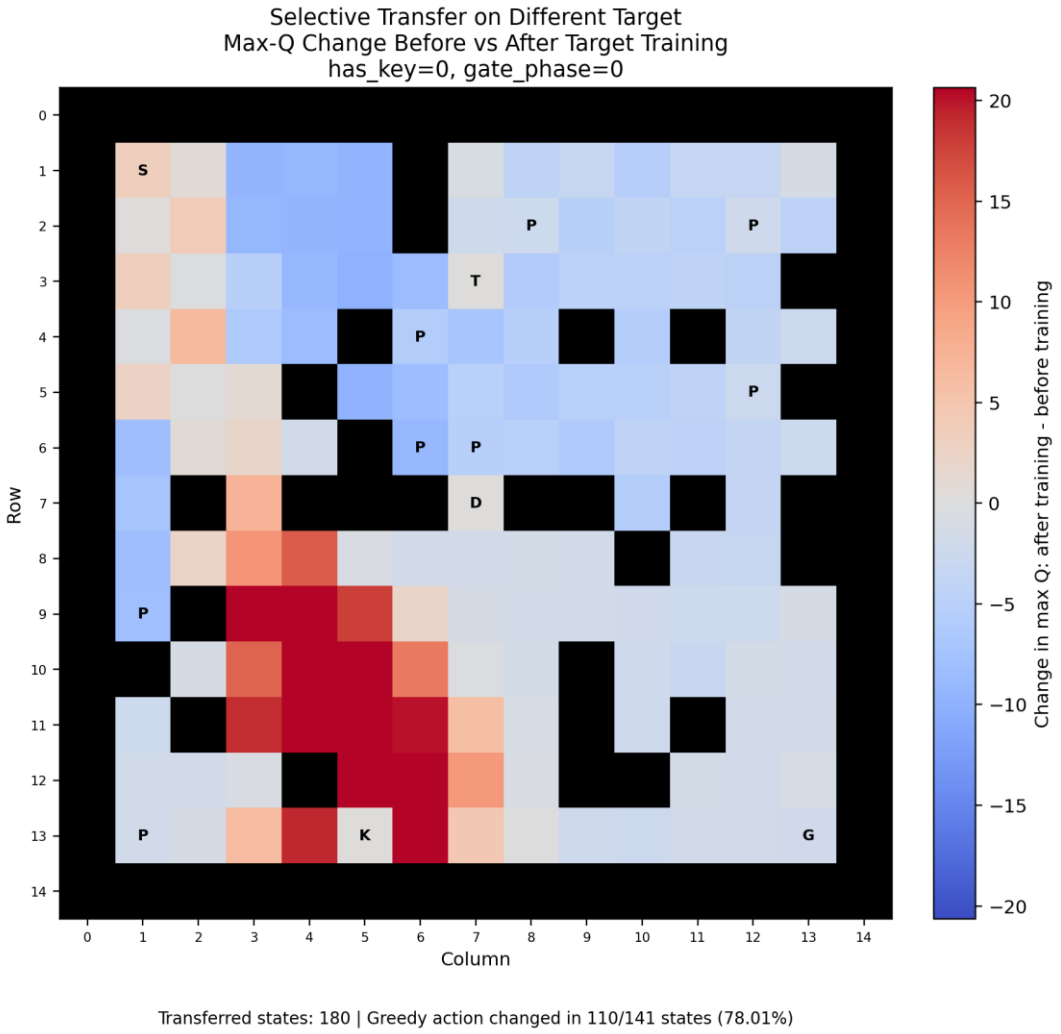
۸،۱،۲. یک به‌روزرسانی واقعی Q در سناریوی انتقال منفی

ویژگی	مقدار
state	[1, 1, 0, 0]
action	RIGHT (3)
reward	-0.9300
next_state	[1, 2, 0, 1]
old_q	-11.804122
next_max_q	-11.056944
target	-11.434097
td_error	0.370026
alpha	0.10
gamma	0.95
new_q	-11.767120

در این نمونه واقعی، مقدار هدف از رابطه

$$target = r + \gamma \max_{a} Q(s', a)$$

از -۱۱.۸۰۴۱۲۲ به -۱۱.۷۶۷۱۲۰ RIGHT برای عمل Q برابر با +۰.۳۷۰۰۲۶ است؛ بنابراین مقدار TD محاسبه شده است. خطای نشان می‌دهد که تجربه جدید، برآورد منفی قبلی را اندکی اصلاح کرده است. تکرار چنین TD افزایش یافته است. مثبت بودن خطای به‌روزرسانی‌هایی در طول آموزش، به‌تدریج باعث جبران اثر انتقال منفی شده است.



شکل ۱۰ ـ نقشه تغییر بیشینه مقدار Q پیش و پس از آموزش مقصد در انتقال Selective به محیط متفاوت

این Heatmap تغییر بیشینه مقدار Q هر حالت را از زمان انتقال اولیه تا پایان آموزش در محیط مقصد متفاوت نشان می‌دهد؛ یعنی برای هر حالت مقدار max Q_after − max Q_before محاسبه شده است. نواحی متمایل به قرمز بیانگر افزایش ارزش برآوردشده پس از تجربه‌کردن محیط مقصد و نواحی متمایل به آبی نشان‌دهنده کاهش ارزش یا اصلاح خوش‌بینی ناشی از دانش منتقل‌شده هستند. خانه‌های سیاه نیز حالت‌های غیرقابل‌عبور یا فاقد مقدار قابل‌مقایسه را نمایش می‌دهند.

در انتقال انتخابی، ۱۸۰ حالت از محیط مبدأ منتقل شدند. پس از آموزش مقصد، عمل حریصانه در ۱۱۰ حالت از ۱۴۱ حالت قابل‌مقایسه تغییر کرد که برابر با ۷۸.۰۱ درصد است. این میزان تغییر بالا نشان می‌دهد حتی پس از حذف حالت‌های دارای همسایگی متفاوت، بخش بزرگی از سیاست منتقل‌شده با ساختار و پاداش‌های مقصد سازگار نبوده و عامل مجبور شده است آن را با تجربه‌های جدید اصلاح کند. در عین حال، رسیدن عملکرد نهایی به موفقیت ۱۰۰ درصد نشان می‌دهد Q-Learning توانسته اثر نامناسب دانش اولیه را به‌تدریج جبران کند.

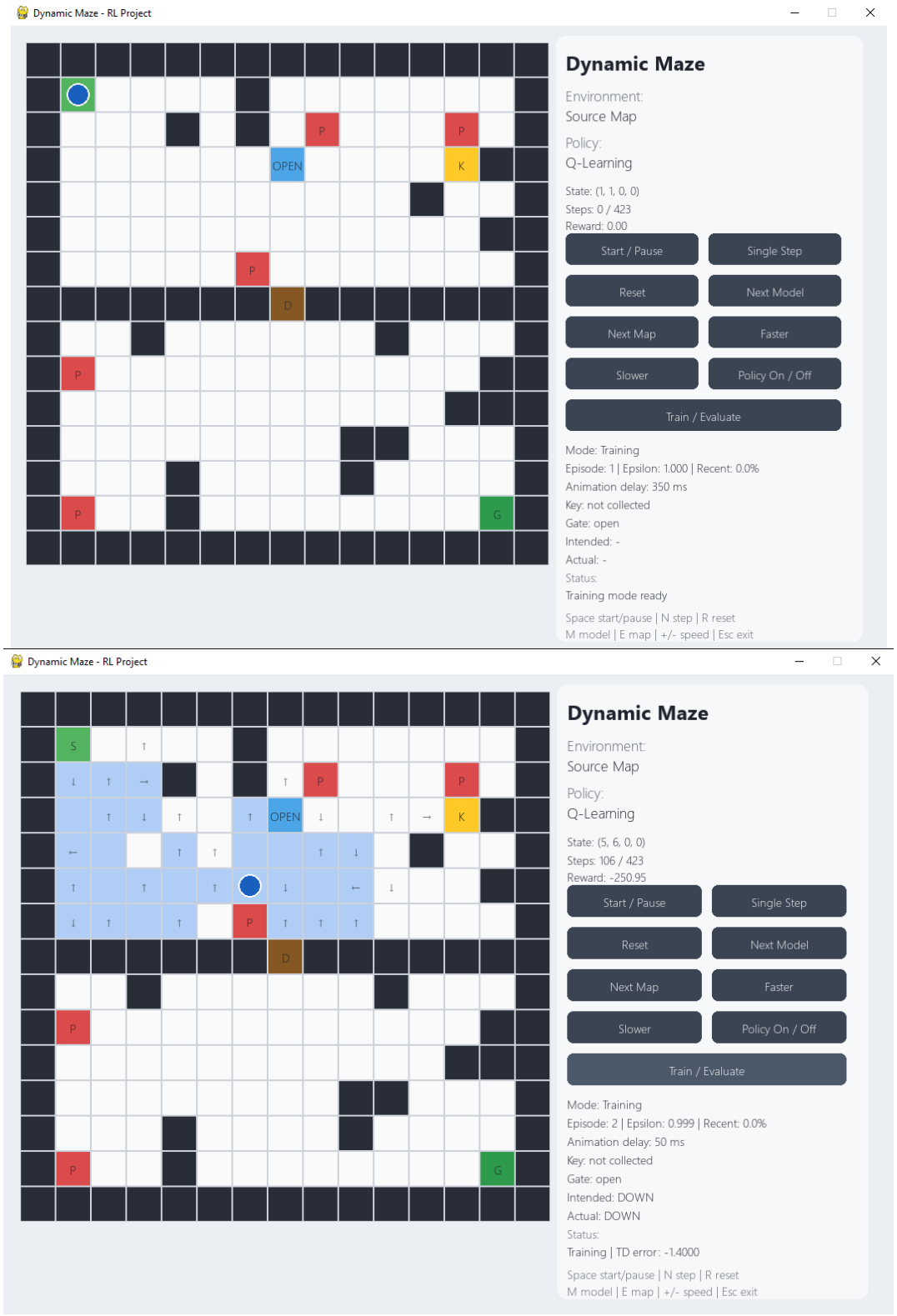
۹. رابط گرافیکی، آزمون‌ها و بازتولیدپذیری

۹,۱. رابط گرافیکی

رابط Pygame امکان انتخاب نقشه و مدل، شروع و توقف، اجرای تک‌گام، Reset، تغییر سرعت و مشاهده مسیر عامل را فراهم می‌کند. وضعیت جاری، تعداد گام، پاداش، وضعیت کلید و دروازه، عمل موردنظر و عمل واقعی در پنل کنار نقشه نمایش داده می‌شوند.

python gui/maze_gui.py

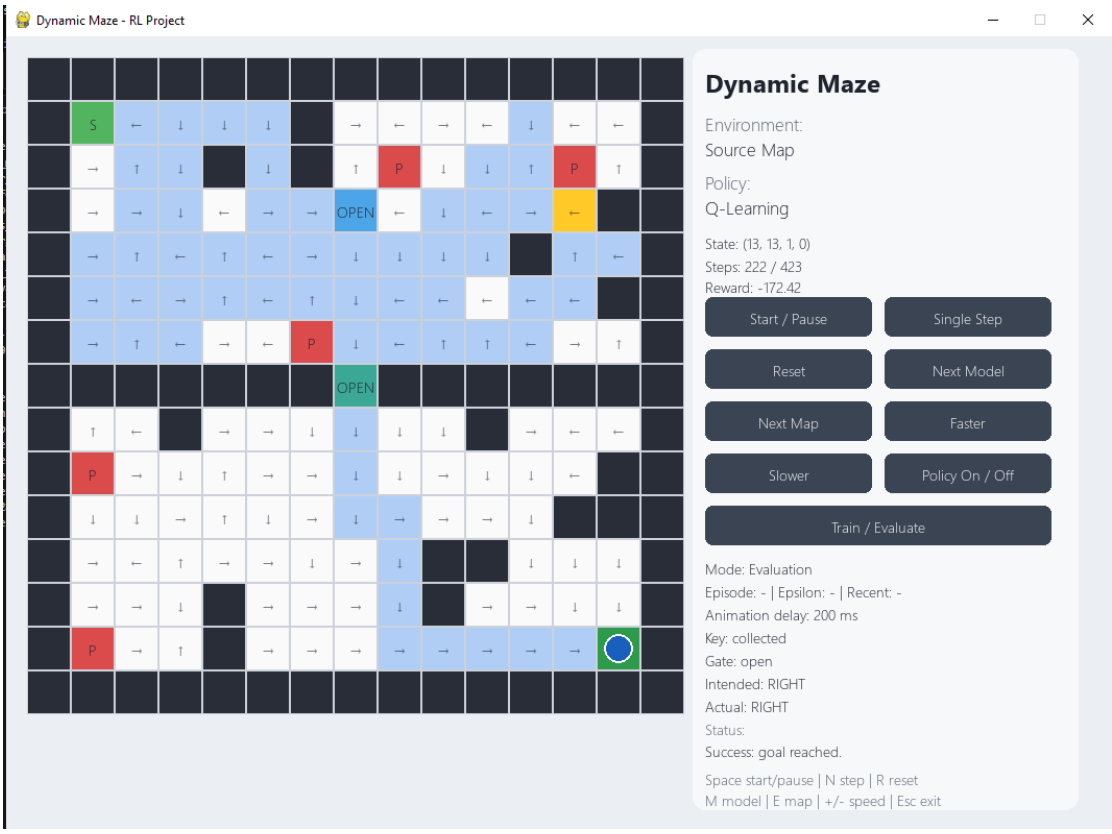
رابط گرافیکی علاوه بر انتخاب نقشه و مدل، امکان جابه‌جایی میان حالت‌های Evaluation و Training را فراهم می‌کند. در حالت Evaluation سیاست ذخیره‌شده اجرا می‌شود و فلش‌های سیاست را می‌توان روشن یا مخفی کرد. در حالت Training، عامل Q-Learning به‌صورت برخط آموزش می‌بیند و شماره اپیزود، مقدار epsilon، نرخ موفقیت اپیزودهای اخیر و خطای TD نمایش داده می‌شوند. کنترل شروع، توقف، اجرای تک‌گام، Reset و سرعت انیمیشن نیز در هر دو حالت در دسترس است.



شکل ۱۱ - رابط گرافیکی در حالت Training و نمایشEpisode ، Epsilon، نرخ موفقیت اخیر وTD error

این تصویر اجرای برخط Q-Learning را در حالت Training نشان می‌دهد. در این حالت، عامل هم‌زمان با تعامل با محیط، جدول Q را به‌روزرسانی می‌کند و فلش‌های سیاست نیز متناسب با مقادیر یادگرفته‌شده تغییر می‌کنند. نمایش شماره اپیزود و مقدار epsilonامکان مشاهده روند کاهش اکتشاف را فراهم می‌سازد؛ نرخ موفقیت اخیر نیز یک معیار کوتاه‌مدت از پیشرفت یادگیری ارائه می‌دهد. مقدار TD error جهت و اندازه اصلاح آخرین مقدار Q را نشان می‌دهد؛ خطای مثبت موجب افزایش و خطای منفی

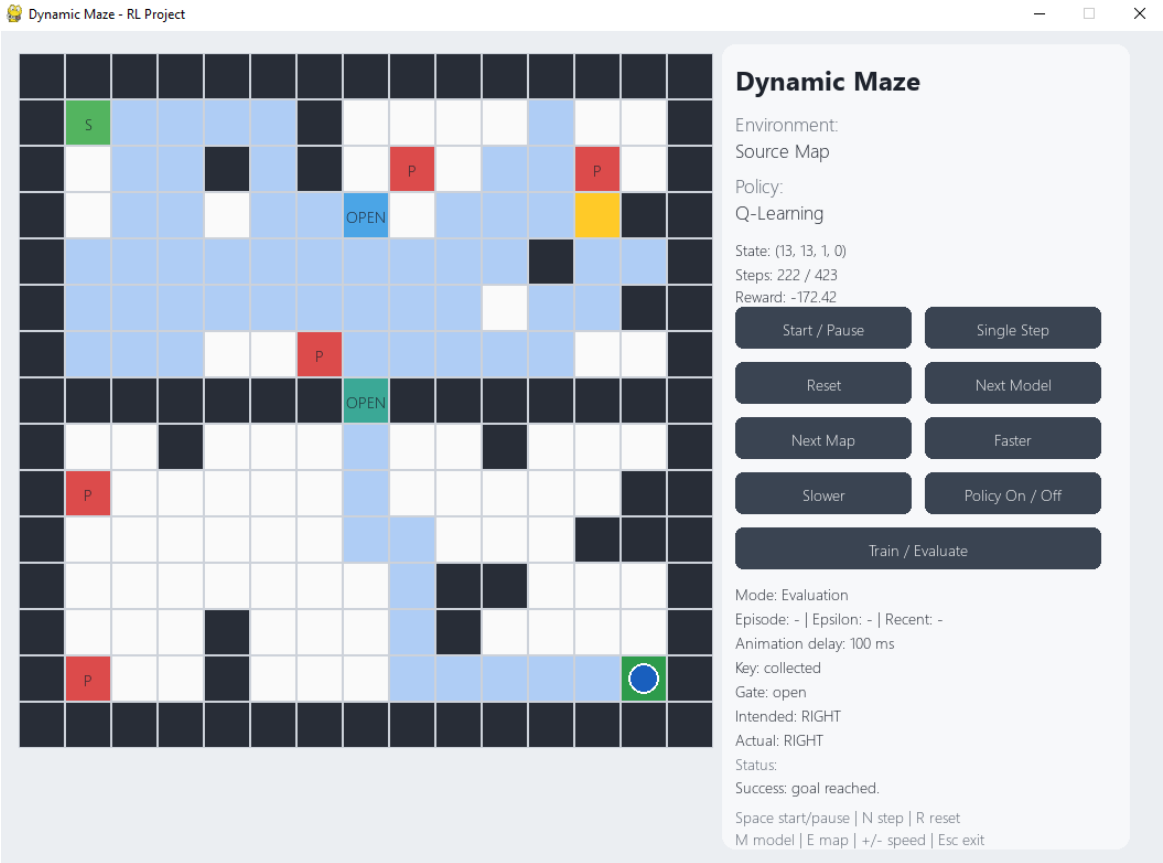
موجب کاهش برآورد زوج حالت—عمل می‌شود. این نمایش برای تحلیل آموزشی مفید است، اما نرخ موفقیت لحظه‌ای به‌تنهایی معیار نهایی کیفیت سیاست نیست و ارزیابی بدون اکتشاف باید جداگانه انجام شود.



شکل ۱۲ - رابط گرافیکی در حالت Evaluation و نمایش سیاست ذخیره‌شده

در حالت Evaluation ، آموزش و اکتشاف متوقف می‌شوند و عامل با استفاده از سیاست حریصانه ذخیره‌شده در محیط حرکت می‌کند. فلش‌های نمایش‌داده‌شده در خانه‌های نقشه، عمل دارای بیشترین مقدار Q را برای وضعیت فعلی کلید و فاز دروازه مشخص می‌کنند و با تغییر این متغیرها ممکن است جهت سیاست نیز تغییر کند. پنل کناری امکان مشاهده تعداد گام، پاداش تجمعی، وضعیت کلید، فاز دروازه، عمل انتخاب‌شده و عمل واقعی اجراشده را فراهم می‌سازد. تفاوت میان عمل انتخاب‌شده و عمل واقعی ناشی از تصادفی‌بودن انتقال محیط است؛ زیرا عمل موردنظر با احتمال ۰.۸ اجرا می‌شود و در سایر موارد ممکن است یکی از اعمال عمود رخ دهد. قابلیت اجرای تک‌گام و تغییر سرعت نیز بررسی مسیر و نقاط تصمیم‌گیری سیاست را آسان‌تر می‌کند.

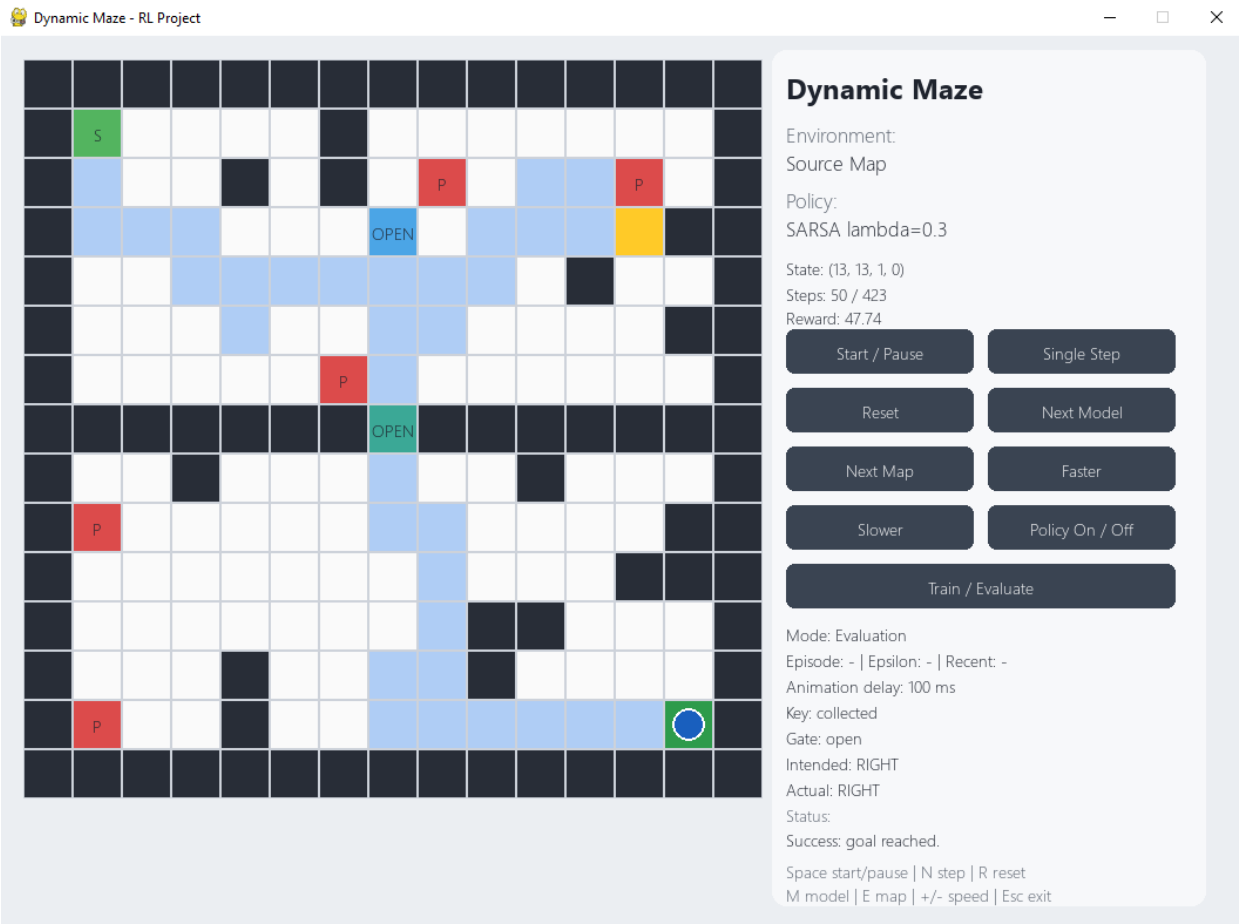
۹,۲. ارزیابی و مقایسه مسیر نهایی سیاست‌های Value Iteration ، Q-Learning و SARSA(λ)



شکل ۱۳ ـ مسیر نهایی عامل Q-Learning در حالت Evaluation روی نقشه اصلی

مسیر نهایی عامل Q-Learning در حالت Evaluation روی نقشه اصلی؛ عامل از نقطه شروع حرکت کرده، کلید را جمع‌آوری کرده، از در عبور کرده و به هدف رسیده است.

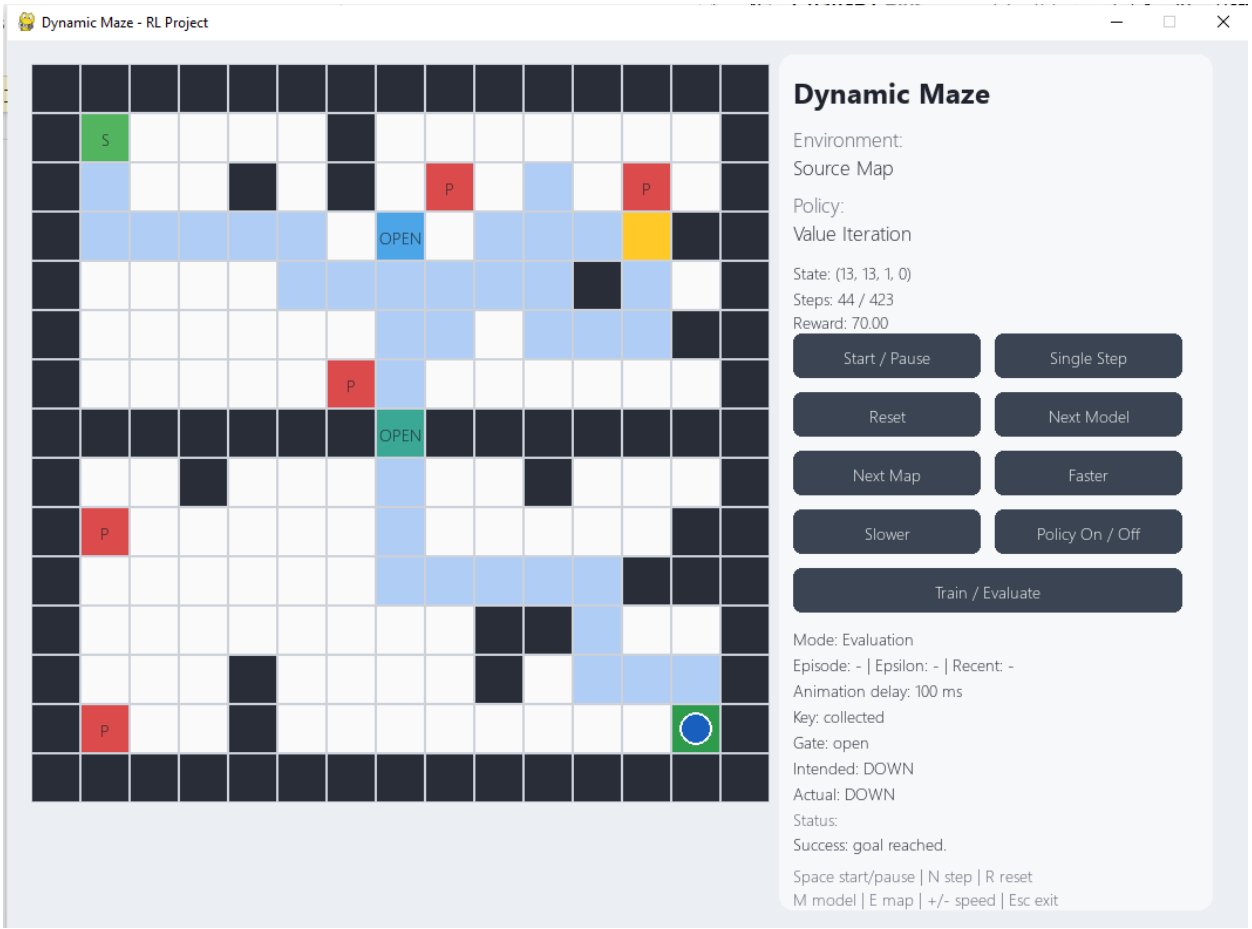
در این شکل، مسیر نهایی عامل در اجرای ارزیابی نهایی Q-Learning نمایش داده شده است. عامل با استفاده از سیاست آموخته‌شده از حالت شروع حرکت می‌کند، ابتدا کلید را جمع‌آوری می‌کند، سپس با باز شدن مسیر یا در، به سمت هدف ادامه می‌دهد و در نهایت به خانه هدف می‌رسد. مشاهده این مسیر نشان می‌دهد که سیاست نهایی فقط از نظر نرخ موفقیت مناسب نیست، بلکه از نظر ترتیب منطقی تصمیم‌ها نیز صحیح عمل می‌کند؛ یعنی ابتدا پیش‌شرط لازم برای رسیدن به هدف (جمع‌آوری کلید) را انجام داده و سپس ادامه مسیر را طی می‌کند.



شکل ۱۴ ـ مسیر نهایی عامل SARSA(λ=0.3) در حالت Evaluation روی نقشه اصلی

در این شکل، یک اجرای نهایی از سیاست آموخته‌شده توسط الگوریتم SARSA(λ=0.3) در حالت Evaluation نمایش داده شده است. در این حالت، فرایند آموزش متوقف است و عامل بدون اکتشاف و بدون تغییر مقادیرQ ، تنها بر اساس سیاست حریصانه ذخیره‌شده حرکت می‌کند. مسیر ثبت‌شده نشان می‌دهد عامل توانسته است ترتیب منطقی مسئله را رعایت کند؛ به این صورت که ابتدا از نقطه شروع به سمت کلید حرکت کرده، پس از جمع‌آوری کلید مسیر مناسب برای عبور از در را انتخاب کرده و در ادامه به خانه هدف رسیده است.

عملکرد صحیح عامل در این مسیر نشان می‌دهد که ردپاهای شایستگی در SARSA(λ) توانسته‌اند اثر پاداش‌ها و جریمه‌های جدید را به چند تصمیم گذشته منتقل کنند و اصلاح مسیر را تنها به آخرین زوج حالت–عمل محدود نکنند. مقدار λ برابر ۰.۳ در آزمایش‌های انجام‌شده بهترین توازن را میان نرخ موفقیت، پاداش، طول مسیر و زمان آموزش ایجاد کرد. این تنظیم در ارزیابی‌های کلی به موفقیت ۱۰۰ درصد، میانگین پاداش ۵۶.۰۲ و میانگین ۴۹.۸۴ گام رسید. البته تصویر حاضر فقط یک اجرای نمونه را نشان می‌دهد و تعداد گام و پاداش نوشته‌شده در پنل GUI مربوط به همان اجرای منفرد است، در حالی که مقادیر جدول نتایج از میانگین چند اجرای ارزیابی به دست آمده‌اند.



شکل ۱۵ ـ مسیر نهایی عامل Value Iteration با γ=0.95 در حالت Evaluation روی نقشه اصلی

این شکل مسیر نهایی عامل را بر اساس سیاست حریصانه استخراج‌شده از الگوریتم Value Iteration با ضریب تنزیل γ=0.95 نشان می‌دهد. برخلاف Q-Learning و SARSA که مقادیر خود را از تجربه‌های نمونه‌برداری‌شده می‌آموزند، Value Iterationاز مدل کامل انتقال و پاداش محیط استفاده می‌کند و ارزش تمام حالت‌های معتبر را به‌صورت نظام‌مند محاسبه می‌کند. در نتیجه، مسیر مشاهده‌شده حاصل اجرای سیاستی است که پس از همگرایی تابع ارزش استخراج شده است.

در اجرای نشان‌داده‌شده، سیاست توانسته است وابستگی چندمرحله‌ای مسئله را در نظر بگیرد؛ یعنی عامل مستقیماً به سمت هدف نرفته، بلکه ابتدا پیش‌شرط لازم برای ادامه مسیر را با جمع‌آوری کلید انجام داده و سپس از در عبور کرده و به هدف رسیده است. این رفتار نشان می‌دهد که نمایش حالت شامل موقعیت، وضعیت کلید و فاز دروازه، اطلاعات لازم برای تصمیم‌گیری صحیح را در اختیار الگوریتم قرار داده است.

مدل γ=0.95 پس از ۳۶۱ تکرار و با مقدار نهایی delta کمتر از 10^{-8}همگرا شد و به دلیل توجه بیشتر به پاداش‌های بلندمدت به‌عنوان سیاست مرجع انتخاب گردید. بنابراین، این تصویر علاوه بر نمایش مسیر نهایی، نمونه‌ای بصری از سیاست مرجع پروژه است که سایر سیاست‌های model-free با آن مقایسه شده‌اند

مقایسه عملکرد سیاست‌های Value Iteration ، Q-Learning و SARSA(λ) با پاداش Sparse

الگوریتم	تنظیم اصلی	تعداد اجرای مستقل	نرخ موفقیت ارزیابی	میانگین پاداش	میانگین تعداد گام
Value Iteration	γ=0.95	1	100.00٪	65.74	47.60
Q-Learning	کاهش خطیε	3	98.67 ± 1.53٪	42.30 ± 4.60	65.46 ± 2.98
SARSA(λ)	λ=0.3	3	100.00 ± 0.00٪	57.46 ± 3.13	54.17 ± 1.83

جدول بالا عملکرد سه سیاست را در شرایط دارای پاداش Sparse مقایسه می‌کند. نتایج Q-Learning و SARSA(λ=0.3) از میانگین سه اجرای مستقل باSeed های ۸، ۱۸ و ۲۸ به دست آمده‌اند و مقدار پس از علامت ± انحراف معیار میان اجراها را نشان می‌دهد. سیاست Value Iteration با γ=0.95 پس از همگرایی در ۳۶۱ تکرار، طی ۱۰۰ اپیزود مستقل ارزیابی شده است. از آنجا که برای Value Iteration فقط یک اجرای ارزیابی ثبت شده، انحراف معیار میان چند اجرای مستقل برای آن گزارش نشده است.

Value Iterationدر ارزیابی انجام‌شده به نرخ موفقیت ۱۰۰ درصد، میانگین پاداش ۶۵.۷۴ و میانگین ۴۷.۶۰ گام رسید و از نظر پاداش و کوتاهی مسیر بهترین نتیجه ثبت‌شده را به دست آورد. این عملکرد با دسترسی الگوریتم به مدل کامل انتقال و پاداش محیط سازگار است؛ زیرا Value Iteration می‌تواند همه حالت‌های معتبر و پیامدهای احتمالی اعمال را به‌صورت نظام‌مند بررسی کند.

SARSA(λ=0.3)نیز در هر سه اجرای مستقل نرخ موفقیت ۱۰۰ درصد داشت و با میانگین پاداش ۵۷.۴۶ و میانگین ۵۴.۱۷ گام، بهترین عملکرد را در میان دو روش model-free نشان داد. انحراف معیار پایین تعداد گام‌ها، برابر ۱.۸۳، بیانگر پایداری مناسب این روش میانSeed های مختلف است. استفاده از ردپای شایستگی باعث شده است اثر پاداش و جریمه به چند تصمیم قبلی منتقل شود و عامل بتواند مسیر خود را با سرعت بیشتری اصلاح کند.

Q-Learningبا نرخ موفقیت میانگین ۹۸.۶۷ درصد همچنان عملکرد مناسبی داشت، اما میانگین پاداش آن ۴۲.۳۰ و میانگین طول مسیر آن ۶۵.۴۶ گام بود. این نتایج نشان می‌دهند که عامل Q-Learning در بعضی اجراها مسیرهای طولانی‌تری طی کرده یا جریمه بیشتری دریافت کرده است. نرخ موفقیت کمتر از ۱۰۰ درصد نیز نشان می‌دهد که در تعداد محدودی از اپیزودهای ارزیابی، عامل پیش از رسیدن به هدف به سقف گام رسیده است.

در مجموع، در آزمایش‌های دارای پاداشSparse ، Value Iterationبهترین عملکرد عددی ثبت‌شده را داشت، SARSA(λ=0.3)بهترین روش model-free بود و Q-Learning در جایگاه بعدی قرار گرفت. بااین‌حال، مقایسه پایداری میان الگوریتم‌ها باید با احتیاط انجام شود؛ زیرا نتایج Q-Learning و SARSA میان سه اجرای مستقل خلاصه شده‌اند، در حالی که ارزیابی Value Iteration فعلاً تنها برای یک اجرای همگرا انجام شده است. برای یک مقایسه آماری کاملاً هم‌سطح، می‌توان در آینده ارزیابی Value Iteration را نیز با چند Seed مستقل تکرار کرد.

مقایسه مسیر نهایی سه سیاست

تصاویر بالا رفتار نهایی سه روش Value Iteration ، Q-Learning و SARSA(λ=0.3) را در یک محیط و در حالت Evaluation نشان می‌دهند. در هر سه اجرا، آموزش و اکتشاف متوقف شده و عامل بر اساس سیاست حریصانه نهایی خود حرکت کرده است. هر سه سیاست توانسته‌اند ترتیب اساسی مسئله را رعایت کنند؛ عامل ابتدا کلید را جمع‌آوری کرده، سپس امکان عبور از در را به دست آورده و در نهایت به هدف رسیده است. این نتیجه نشان می‌دهد که هر سه روش، با وجود تفاوت در شیوه یادگیری یا محاسبه سیاست، ساختار چندمرحله‌ای مسئله را به‌درستی در تصمیم‌های خود منعکس کرده‌اند.

بااین‌حال، رسیدن هر سه عامل به هدف به معنای یکسان‌بودن مسیرها و سیاست‌های آن‌ها نیست Value Iteration .از مدل کامل انتقال و پاداش محیط استفاده می‌کند و تمام حالت‌های معتبر را به‌صورت نظام‌مند بررسی می‌کند. در نتیجه، سیاست استخراج‌شده از آن در تنظیمات ثابت، یک سیاست مرجع و مستقل از مسیرهای تصادفی آموزش است؛ هرچند اجرای این سیاست در محیط تصادفی می‌تواند در اپیزودهای مختلف مسیرهای متفاوتی ایجاد کند Q-Learning .یک روش model-free و off-policy است و مقادیر Q را با استفاده از بیشینه مقدار حالت بعدی به‌روزرسانی می‌کند. در مقابل، SARSA یک روش on-policy است و در به‌روزرسانی خود عمل واقعاً انتخاب‌شده توسط سیاست رفتاری را در نظر می‌گیرد. همچنین استفاده از ردپای شایستگی در SARSA(λ) باعث می‌شود اثر یک پاداش یا جریمه به چند تصمیم قبلی نیز منتقل شود.

برای مقایسه کمی منصفانه، نتایج هر سه الگوریتم با پاداش Sparse بررسی شدند. سیاست Value Iteration با γ=0.95 پس از همگرایی در ۳۶۱ تکرار، در ۱۰۰ اپیزود مستقل ارزیابی شد و نرخ موفقیت ۱۰۰ درصد، میانگین پاداش ۶۵.۷۴ و میانگین ۴۷.۶۰ گام را ثبت کرد Q-Learning .با برنامه کاهش خطی epsilon ، بر اساس میانگین سه اجرای مستقل با Seed های ۸، ۱۸ و ۲۸، به نرخ موفقیت 98.67 ± 1.53 درصد، میانگین پاداش 42.30 ± 4.60 و میانگین 65.46 ± 2.98 گام رسید SARSA .با λ=0.3 نیز در سه اجرای مستقل نرخ موفقیت 100.00 ± 0.00 درصد، میانگین پاداش 57.46 ± 3.13 و میانگین 54.17 ± 1.83 گام را به دست آورد.

این نتایج نشان می‌دهند که Value Iteration در ارزیابی انجام‌شده از نظر میانگین پاداش و کوتاهی مسیر، بهترین عملکرد عددی را داشته است. این برتری با دسترسی آن به مدل کامل محیط و بررسی نظام‌مند حالت‌ها سازگار است. در میان الگوریتم‌های model-free، SARSA(λ=0.3) نسبت به Q-Learning عملکرد بهتری داشت؛ زیرا علاوه بر موفقیت کامل در هر سه اجرای مستقل، میانگین پاداش بیشتر و تعداد گام کمتری ثبت کرد. انحراف معیار پایین تعداد گام‌های SARSA نیز نشان‌دهنده پایداری مناسب آن میان Seed های مختلف است.

تفاوت مسیرهای مشاهده‌شده در تصاویر می‌تواند علاوه بر تفاوت الگوریتم‌ها، ناشی از تصادفی‌بودن انتقال‌های محیط باشد. در هر گام، عمل انتخابی با احتمال ۰.۸ اجرا می‌شود و با احتمال ۰.۱ ممکن است یکی از دو حرکت عمود رخ دهد. بنابراین، حتی یک سیاست ثابت نیز در دو اجرای جداگانه الزاماً مسیر کاملاً یکسانی طی نمی‌کند. ازاین‌رو، تصاویر ارائه‌شده نمونه‌های بصری و قابل تفسیر از رفتار سیاست‌ها هستند و نتیجه‌گیری اصلی باید بر میانگین نرخ موفقیت، پاداش و تعداد گام در مجموعه اپیزودهای ارزیابی متکی باشد.

در مجموع، Value Iteration یک سیاست مرجع مبتنی بر مدل کامل فراهم کرد و در ارزیابی Sparse بهترین نتیجه عددی را به دست آورد Q-Learning .نشان داد که بدون دسترسی به مدل نیز می‌توان سیاستی با نرخ موفقیت بالا آموخت SARSA(λ=0.3) . نیز بهترین روش model-free در آزمایش‌های Sparse بود و با درنظرگرفتن عمل واقعی سیاست و انتقال خطای TD به تصمیم‌های گذشته، مسیرهای کوتاه‌تر و پاداش بیشتری نسبت به Q-Learning ایجاد کرد. نمایش تصویری هر سه مسیر، نتایج عددی

گزارش را تکمیل می‌کند و نشان می‌دهد موفقیت الگوریتم‌ها علاوه بر فایل‌های آماری، در اجرای واقعی رابط گرافیکی نیز قابل مشاهده است.

۹,۳. آزمون‌های واحد

مجموعه آزمون‌ها شامل رفتار محیط، احتمال انتقال، برخورد با دیوار، جمع‌آوری کلید، در قفل، دروازه دوره‌ای، همگرایی Value Iteration، به‌روزرسانی Q ، برنامه‌های epsilon ، ردپای SARSA و انتقال یادگیری است. آخرین اجرای ثبت‌شده شامل ۲۹ آزمون موفق است.

۹,٤. بازتولید نتایج

```
python -m pip install -r requirements.txt
python environments/generator.py
python experiments/run_value_iteration.py
python experiments/run_q_learning.py
python experiments/run_sarsa_lambda.py
python experiments/run_reward_comparison.py
python transfer/transfer_learning.py
python experiments/run_transfer_learning.py
python analysis/generate_figures.py
python analysis/generate_reward_comparison_figures.py
python analysis/generate_transfer_q_change.py
python -m unittest discover -s tests -v
python experiments/evaluate_value_iteration.py
python analysis/generate_value_iteration_convergence.py
python analysis/generate_q_learning_curve.py
python analysis/generate_sarsa_lambda_curves.py
```

```
----------------------------------------------------------------------
Ran 29 tests in 4.235s

OK
PS C:\Users\pc\Documents\RL_FinalProject_40403384>
```

تمام مسیرها نسبت به ریشه پروژه محاسبه می‌شوند. Seed های ثابت برای تولید نقشه، انتقال‌های محیط، اکتشاف عامل، ارزیابی و ساخت نقشه‌های مقصد استفاده شده‌اند. نمودارها مستقیماً از فایل‌های خام مخزن تولید می‌شوند.

اسکریپت run_reward_comparison.py مقایسه چند Seed میان پاداش‌های Sparse و Shaped را اجرا می‌کند.

اسکریپت generate_reward_comparison_figures.py نمودارهای سرعت یادگیری، کیفیت مسیر و توافق سیاست را

تولید می‌کند و اسکریپت generate_transfer_q_change.py تغییر بیشینه مقدار Q و تغییر عمل حریصانه پیش و پس از آموزش مقصد را در سناریوی انتقال انتخابی نمایش می‌دهد.

۱۰. محدودیت‌ها و پیشنهادهای بهبود

• توافق سیاست Q-Learning با سیاست مرجع Value Iteration در آزمایش Sparse و Seed پایه ۸ برابر ۵۵.۳۲ درصد بود. این مقدار نشان می‌دهد موفقیت بالای یک سیاست الزاماً به معنای یکسان‌بودن عمل حریصانه آن با سیاست مرجع در تمام حالت‌ها نیست. افزایش تعداد اپیزودهای آموزش، بررسی حالت‌های کم‌بازدید و تنظیم دقیق‌تر نرخ یادگیری می‌تواند علت این اختلاف را روشن‌تر کند.

• برخی نتایج انتقال در محیط متفاوت نشان‌دهنده انتقال منفی اولیه هستند؛ معیار شباهت دقیق‌تر یا انتقال مبتنی بر اطمینان می‌تواند مفید باشد.

• پاداش شکل‌داده‌شده سرعت یادگیری را بالا می‌برد، اما باید بررسی شود که سیاست بهینه مسئله اصلی را تغییر ندهد.

• مقایسه Sparse و Shaped در این پروژه با سه Seed مستقل ۸، ۱۸ و ۲۸ انجام شده است؛ بااین‌حال برای برآورد آماری مطمئن‌تر، بهتر است تعداد اجراها در مطالعات بعدی به دست‌کم ۱۰ Seed افزایش یابد و علاوه بر میانگین و انحراف معیار، فاصله اطمینان ۹۵ درصد نیز گزارش شود.

• اندازه‌گیری حافظه و پایداری میان چند اجرای مستقل می‌تواند در نسخه بعدی دقیق‌تر و با نمودار فاصله اطمینان گزارش شود.

• ظاهر نمودار سیاست در خانه‌های ویژه می‌تواند با حذف فلش روی S/K/D/G/P و افزودن legend خواناتر شود.

۱۱. جمع‌بندی

11-1. جمع‌بندی کلی پروژه

در این پروژه، یک چارچوب کامل و قابل بازتولید برای حل مسئله هزارتوی پویا با استفاده از روش‌های یادگیری تقویتی طراحی و پیاده‌سازی شد. مسئله فقط یافتن یک مسیر ثابت از نقطه شروع به هدف نبود، بلکه عامل باید در محیطی تصادفی و دارای وابستگی زمانی حرکت می‌کرد، کلید را پیش از عبور از در قفل جمع‌آوری می‌کرد، وضعیت باز یا بسته‌بودن دروازه دوره‌ای را در نظر می‌گرفت، از خانه‌های جریمه دوری می‌کرد و با وجود احتمال لغزش در اجرای اعمال، سیاستی پایدار برای رسیدن به هدف می‌آموخت. به همین دلیل، حالت محیط به‌صورت چهار مؤلفه‌ای (row, column, has_key, gate_phase) تعریف شد تا علاوه بر مکان عامل، وضعیت کلید و فاز زمانی دروازه نیز در تصمیم‌گیری وارد شوند. این تعریف سبب شد محیط از دید عامل دارای خاصیت مارکوف باشد و مقایسه میان الگوریتم‌های مختلف بر مبنای یک نمایش حالت مشترک انجام شود. پروژه شامل سه روش اصلی Value Iteration ،

Q-LearningوSARSA(λ) ، دو نوع پاداش Sparse وShaped ، چند سناریوی انتقال یادگیری، رابط گرافیکیPygame ، آزمون‌های واحد، مدل‌های ذخیره‌شده، فایل‌های داده خام و نمودارهای تحلیلی قابل تولید مجدد است .

یکی از دستاوردهای اصلی پروژه، اتصال مستقیم گزارش به خروجی واقعی برنامه است. نتایج عددی از فایل‌های CSV و JSON خوانده شده‌اند، مدل‌های نهایی در قالب JSON ذخیره شده‌اند و شکل‌ها نیز از داده‌های خام تولید شده‌اند. علاوه بر گزارش معیارهای کلی، نمونه‌های واقعی به‌روزرسانیQ ، خطایTD ، ردپای شایستگی و اصلاح مقادیر Q پس از انتقال منفی نیز استخراج و تحلیل شدند. بنابراین، نتیجه‌گیری‌های ارائه‌شده فقط توصیف نظری الگوریتم‌ها نیستند، بلکه به اجرای واقعی، داده‌های ثبت‌شده و رفتار قابل مشاهده عامل متصل‌اند.

11-2. نتیجه‌گیری مربوط بهValue Iteration

در بخشValue Iteration ، سه مقدار ضریب تنزیل γ=0.80، γ=0.90و γ=0.95بررسی شدند. هر سه اجرا با آستانه همگرایی 10^{-8}با موفقیت همگرا شدند، اما افزایش γ باعث افزایش محسوس تعداد تکرارها و زمان محاسبه شد. الگوریتم برای γ=0.80پس از ۸۴ تکرار، برای γ=0.90پس از ۱۷۶ تکرار و برای γ=0.95پس از ۳۶۱ تکرار به آستانه همگرایی رسید. مقدار نهایی delta در هر سه حالت کمتر از 10^{-8}بود و این موضوع نشان می‌دهد فرایند تکرار مقدار با معیار توقف تعیین‌شده به‌درستی خاتمه یافته است. افزایش تعداد تکرارها درγ های بزرگ‌تر قابل انتظار است، زیرا در این حالت پاداش‌های دورتر وزن بیشتری دارند و اطلاعات ارزش باید در فاصله بیشتری از هدف، کلید و سایر نقاط مهم محیط منتشر شود. نتایج عددی اولیه گزارش نیز نشان می‌دادند که تعداد تکرار برایγ های ۰.۸۰، ۰.۹۰ و ۰.۹۵ به‌ترتیب ۸۴، ۱۷۶ و ۳۶۱ است .

مدل γ=0.95به‌عنوان مرجع اصلی انتخاب شد. دلیل این انتخاب فقط داشتن بیشترین γ نبود؛ بلکه ساختار مسئله شامل چند تصمیم وابسته به آینده است. عامل باید گاهی از مسیر مستقیم هدف دور شود تا کلید را جمع‌آوری کند، سپس به در قفل برسد و در زمان مناسب از دروازه دوره‌ای عبور کند. در چنین محیطی، داشتن افق بلندتر باعث می‌شود پاداش نهایی هدف و اثر تصمیم‌های چندمرحله‌ای با وزن بیشتری در ارزش حالت‌ها منعکس شود. در مقابل، γکوچک‌تر سریع‌تر همگرا شد و هزینه محاسباتی کمتری داشت، اما سیاست آن با سیاست مرجع کاملاً یکسان نبود. میزان توافق سیاستγ های ۰.۸۰ و ۰.۹۰ با γ برابر ۰.۹۵ به‌ترتیب حدود ۹۱.۴۶ و ۹۸.۲۲ درصد بود. این نتیجه نشان می‌دهد γ برابر ۰.۹۰ نیز سیاستی بسیار نزدیک به مرجع ایجاد کرده است، ولی γ برابر ۰.۹۵ برای ادامه تحلیل‌ها انتخاب محافظه‌کارانه‌تری نسبت به پاداش‌های بلندمدت بود .

Heatmapتابع ارزش نشان داد که ارزش حالت‌ها فقط تابع فاصله هندسی تا هدف نیست. خانه‌های نزدیک به کلید، در قفل، مسیرهای عبورپذیر و گلوگاه‌ها ساختار ارزش متفاوتی دارند و دیوارها یا موانع موجب می‌شوند دو خانه با فاصله منهتن یکسان، ارزش‌های متفاوتی داشته باشند. سیاست حریصانه نیز نشان داد که عامل در برخی نقاط ابتدا به سمت کلید هدایت می‌شود و تنها پس از تغییر متغیر has_keyمسیر مناسب به سمت در و هدف را انتخاب می‌کند. بنابراین، Value Iterationتوانست ساختار کاملMDP ، وابستگی به کلید و تغییرات زمانی دروازه را در سیاست خود منعکس کند.

11-3. نتیجه‌گیری مربوط به Q-Learning

درQ-Learning ، دو برنامه کاهش epsilon شامل کاهش خطی و کاهش نمایی بررسی شدند. هر دو برنامه در ۱۰۰ اپیزود پایانی آموزش و در ارزیابی حریصانه به نرخ موفقیت ۱۰۰ درصد رسیدند. بااین‌حال، رسیدن به موفقیت یکسان به معنای یکسان‌بودن کیفیت

سیاست‌ها نبود. در ارزیابی ثبت‌شده، برنامه خطی میانگین پاداش ۹.۷۰ و میانگین ۷۸.۶۷ گام داشت، در حالی که برنامه نمایی میانگین پاداش ۳۲.۵۴- و میانگین ۸۳.۶۵ گام به دست آورد. این تفاوت نشان می‌دهد کاهش خطی epsilon تعادل مناسب‌تری میان اکتشاف اولیه و بهره‌برداری نهایی ایجاد کرده است. هر دو عامل توانسته‌اند به هدف برسند، اما عامل دارای کاهش خطی، در مجموع برخوردهای نامطلوب، جریمه‌ها یا اتلاف گام کمتری داشته است .

نمودار جدید روند آموزش نیز این نتیجه را تأیید کرد. میانگین متحرک نرخ موفقیت در پایان آموزش برای هر دو برنامه به ۱۰۰ درصد رسید، ولی میانگین متحرک پاداش ۵۰ اپیزود پایانی برای کاهش خطی برابر ۳۱.۸۲ و برای کاهش نمایی برابر ۲۶.۹۴ بود. بنابراین، تفاوت دو برنامه فقط در ارزیابی نهایی دیده نشد، بلکه در بخش پایانی روند آموزش نیز برنامه خطی بازده تجمعی بهتری داشت. انتخاب برنامه خطی به‌عنوان تنظیم اصلی Q-Learning بر مبنای ترکیبی از نرخ موفقیت، پاداش، طول مسیر و روند آموزش انجام شد، نه بر اساس یک معیار منفرد.

نقشه فراوانی بازدید حالت‌ها نشان داد که تجربه عامل در تمام فضای حالت یکنواخت نیست. بیشترین بازدیدها در مسیرهای منتهی به کلید، در، گلوگاه‌ها و هدف متمرکز بودند و بسیاری از خانه‌های خارج از مسیرهای رایج کمتر نمونه‌برداری شدند. این مشاهده توضیح می‌دهد که چرا یک الگوریتم model-free ممکن است حتی پس از رسیدن به موفقیت بالا، در برخی حالت‌های کم‌تکرار مقادیر Q دقیق یا سیاستی مشابه مدل برنامه‌ریزی پویا نداشته باشد Q-Learning .برای یادگیری به نمونه‌های واقعی نیاز دارد و حالت‌هایی که کمتر دیده می‌شوند، فرصت کمتری برای اصلاح Q دریافت می‌کنند.

بازسازی عددی یک به‌روزرسانی واقعی نیز صحت پیاده‌سازی را نشان داد. در نمونه ثبت‌شده، عامل در حالت (1,1,0,0)عمل RIGHTرا انتخاب کرد، پاداش تقریبی 0.96-دریافت شد، مقدار Q قبلی صفر بود و چون بیشینه Q حالت بعدی نیز صفر بود، targetو خطای TD برابر 0.96-شدند. با نرخ یادگیری α=0.1، مقدار جدید Q برابر 0.096-به دست آمد. این نمونه نشان می‌دهد عامل حتی در گام‌های اولیه از پیامدهای منفی حرکت یاد می‌گیرد و مقادیر Q به‌تدریج از تجربه واقعی شکل می‌گیرند .

11-4. نتیجه‌گیری مربوط به SARSA(λ)

در SARSA(λ) ، مقادیر λ=0.0، λ=0.3، λ=0.7 و λ=0.9 مقایسه شدند. تمام تنظیم‌ها در ۱۰۰ اپیزود پایانی آموزش به موفقیت ۱۰۰ درصد رسیدند، اما نتایج ارزیابی تفاوت مهمی میان آن‌ها نشان داد. برای λ=0.0 نرخ موفقیت ارزیابی ۹۹ درصد، میانگین پاداش 6.66-و میانگین گام ۹۲.۶۱ بود. برای λ=0.3نرخ موفقیت ۱۰۰ درصد، پاداش ۵۶.۰۲ و میانگین گام ۴۹.۸۴ ثبت شد . λ=0.7 نیز با موفقیت ۱۰۰ درصد، پاداش ۵۳.۴۵ و ۵۱.۶۸ گام عملکرد مناسبی داشت، ولی اندکی ضعیف‌تر از λ برابر ۰.۳ بود. در λ=0.9 نرخ موفقیت به ۹۸ درصد کاهش یافت و با وجود پاداش ۳۵.۵۸، میانگین طول مسیر به ۶۴.۲۹ گام رسید .

منحنی‌های آموزش SARSA نیز تفاوت میان تنظیم‌ها را آشکارتر کردند. هر چهار مقدار λ در پایان به نرخ موفقیت متحرک ۱۰۰ درصد رسیدند، اما میانگین متحرک پاداش نهایی برایλ های ۰.۰، ۰.۳، ۰.۷ و ۰.۹ به‌ترتیب ۳۱.۸۳، ۴۹.۳۰، ۴۸.۳۰ و ۳۸.۵۹ بود. بنابراین، λ=0.3 بهترین پاداش نهایی را ایجاد کرد و λ=0.7 نیز نتیجه‌ای بسیار نزدیک داشت. این نتیجه نشان می‌دهد مقدار متوسط λ می‌تواند اعتبار خطای TD را به چند گام قبلی منتقل کند و سرعت اصلاح مسیر را بالا ببرد، بدون آنکه مانند λ بسیار بزرگ، عامل را بیش‌ازحد به دنباله‌های طولانی و تصادفی حساس کند.

بررسی ردپای شایستگی در یک اپیزود واقعی نشان داد که خطای TD فقط Q زوج حالت–عمل جاری را تغییر نمی‌دهد، بلکه با وزن‌های کاهشی به زوج‌های قبلی نیز منتقل می‌شود. در پیاده‌سازی replacing trace ، ردپای زوج جاری به یک تنظیم شده و سپس تمام ردپاها با ضریب $\gamma\lambda$ کاهش می‌یابند. در نتیجه، یک پاداش یا جریمه جدید می‌تواند چند تصمیم قبلی را نیز اصلاح کند. این مکانیزم، دلیل اصلی برتری SARSA با λ متوسط نسبت به حالت یک‌مرحله‌ای در این محیط است. در عین حال، افزایش λ هزینه محاسباتی را بیشتر کرد؛ زمان آموزش از حدود ۰.۹۰ ثانیه در λ صفر به ۴.۶۴ ثانیه در λ برابر ۰.۹ افزایش یافت. بنابراین، λبرابر ۰.۳ بهترین توازن میان کیفیت سیاست، سرعت یادگیری، طول مسیر و هزینه محاسباتی را فراهم کرد .

11-5. مقایسه الگوریتم‌های model-based وmodel-free

مقایسه Value Iteration با Q-Learning و SARSA نشان داد که موفقیت نهایی بالا الزاماً به معنای یکسان‌بودن سیاست‌ها نیست Value Iteration .از مدل کامل انتقال و پاداش استفاده می‌کند و در هر تکرار همه حالت‌های معتبر را به‌صورت نظام‌مند به‌روزرسانی می‌کند. در مقابل، Q-Learningو SARSA از تجربه‌های نمونه‌برداری‌شده استفاده می‌کنند و کیفیت مقادیر آن‌ها به تعداد بازدید، مسیر اکتشاف، تصادفی‌بودن انتقال‌ها و بودجه محدود آموزش وابسته است. به همین دلیل، در برخی حالت‌ها مخصوصاً حالت‌های کم بازدید، نزدیک دیوارها، گلوگاه‌ها یا خانه‌های دارای چند عمل تقریباً هم‌ارزش، عمل حریصانه الگوریتم‌های model-free با سیاست مرجع متفاوت بود.

در نسخه اصلاح‌شده گزارش، توافق سیاست Q-Learning با سیاست مرجع Value Iteration حدود ۵۵.۳۲ درصد محاسبه شد. این مقدار نباید به‌تنهایی به‌عنوان شکست Q-Learning تفسیر شود، زیرا سیاست Q-Learning در ارزیابی به موفقیت ۱۰۰ درصد رسیده بود. بخشی از اختلاف می‌تواند ناشی از وجود چند عمل تقریباً هم‌ارزش باشد؛ در این حالت، تغییر کوچک در Q باعث تغییر argmaxمی‌شود، بدون آنکه تفاوت بزرگی در کیفیت مسیر ایجاد شود. بخش دیگر اختلاف به نمونه‌برداری ناکافی برخی حالت‌ها و محدودبودن آموزش به ۶۰۰ اپیزود مربوط است. بنابراین، برای مقایسه منصفانه سیاست‌ها باید علاوه بر درصد توافق دقیق اعمال، نرخ موفقیت، پاداش، طول مسیر و اختلاف مقادیر Q نیز در نظر گرفته شوند.

مقایسه Q-Learning و SARSA نیز تفاوت میان رویکرد off-policy و on-policy را نشان داد Q-Learning .مقدار عمل حریصانه حالت بعدی را در target استفاده می‌کند، حتی زمانی که سیاست رفتاری به‌علت epsilon یا تصادفی‌بودن محیط عمل دیگری اجرا کند SARSA target .را بر اساس عمل واقعاً انتخاب‌شده در حالت بعدی می‌سازد و بنابراین مستقیماً رفتار سیاست اکتشافی را یاد می‌گیرد. در محیطی دارای خانه‌های جریمه، لغزش تصادفی و دروازه زمان‌دار، این تفاوت می‌تواند باعث شود SARSA سیاست محتاطانه‌تر و مسیرهای پایدارتر را ترجیح دهد. نتیجه مناسب SARSA با $\lambda=0.3$و میانگین ۴۹.۸۴ گام در ارزیابی نیز با این تفسیر سازگار است.

11-6. نتیجه‌گیری درباره پاداش Sparse وShaped

آزمایش مقایسه پاداش Sparse و Shaped با سه Seed مستقل ۸، ۱۸ و ۲۸ نشان داد که شکل‌دهی پاداش می‌تواند سرعت یادگیری را افزایش دهد، ولی اثر آن بر کیفیت مسیر در همه الگوریتم‌ها یکسان نیست. درQ-Learning ، نسخه Sparse به‌طور متوسط در اپیزود ۳۸۰.۳۳ به موفقیت ۸۰ درصد و در اپیزود ۴۳۷.۶۷ به موفقیت ۹۵ درصد رسید. نسخه Shaped این آستانه‌ها را زودتر و به‌ترتیب در اپیزودهای ۳۳۲.۳۳ و ۴۱۱.۰۰ به دست آورد. بنابراین، شکل‌دهی پاداش در Q-Learning بازخورد جهت‌دارتری

ایجاد کرد و سرعت یادگیری را افزایش داد. بااین‌حال، میانگین گام ارزیابی از ۶۵.۴۶ در Sparse به ۷۷.۱۰ در Shaped افزایش یافت و پراکندگی آن نیز بیشتر شد. این نتیجه نشان می‌دهد یادگیری سریع‌تر لزوماً به معنی کوتاه‌ترین مسیر نهایی نیست .

در SARSA با $\lambda=0.3$، پاداش Shaped هم سرعت یادگیری و هم کیفیت مسیر را بهبود داد. زمان رسیدن به موفقیت ۹۵ درصد از میانگین ۴۵۸ اپیزود در Sparse به ۳۸۷ اپیزود در Shaped کاهش یافت و میانگین گام ارزیابی نیز از ۵۴.۱۷ به ۵۱.۶۳ رسید. هر دو حالت در ارزیابی موفقیت ۱۰۰ درصد داشتند. این نتیجه نشان می‌دهد اثر پاداش شکل‌داده‌شده به الگوریتم وابسته است و باید هم‌زمان با معیارهای موفقیت، طول مسیر و پایداری ارزیابی شود .

توافق سیاست‌های Sparse و Shaped برای Q-Learning حدود ۶۹.۰۴ درصد و برای SARSA حدود ۶۹.۹۳ درصد بود. این اختلاف نشان می‌دهد در بودجه محدود ۶۰۰ اپیزودی، شکل‌دهی پاداش فقط سرعت همگرایی را تغییر نداده، بلکه سیاست حریصانه نهایی را نیز تحت تأثیر قرار داده است. بااین‌حال، از این آزمایش نمی‌توان نتیجه گرفت که شکل‌دهی پاداش حتماً سیاست بهینه نظری مسئله اصلی را تغییر داده است؛ زیرا اختلاف مشاهده‌شده می‌تواند ناشی از آموزش محدود، اکتشاف تصادفی و همگرا نشدن کاملQ ها باشد. همچنین پاداش عددی Sparse و Shaped مستقیماً قابل مقایسه نیست، زیرا مقیاس و تعریف آن‌ها متفاوت است .

11-7. نتیجه‌گیری درباره انتقال یادگیری

بخش انتقال یادگیری نشان داد که سودمندی دانش مبدأ به میزان شباهت ساختاری میان محیط مبدأ و مقصد وابسته است. در نقشه مشابه، انتقال کامل و انتقال‌های مقیاس‌شده عملکرد اولیه بسیار بالایی داشتند. آموزش از صفر فقط ۲ درصد موفقیت اولیه و ۴ درصد موفقیت در ۱۰۰ اپیزود اول داشت و عملکرد نهایی آن نیز ۹۵ درصد بود. در مقابل، انتقال کامل با موفقیت اولیه ۹۸ درصد شروع کرد و پس از آموزش به موفقیت نهایی ۱۰۰ درصد، پاداش ۵۸.۰۱ و میانگین ۴۸.۳۳ گام رسید. انتقال‌های مقیاس‌شده با ضرایب ۰.۵۰ و ۰.۷۵ نیز عملکرد نهایی نزدیک به انتقال کامل داشتند. انتقال انتخابی در نقشه مشابه با موفقیت اولیه ۹۱ درصد و موفقیت نهایی ۱۰۰ درصد نشان داد که حتی انتقال بخشی از حالت‌های معتبر نیز می‌تواند شروع بسیار بهتری نسبت به آموزش از صفر ایجاد کند .

در نقشه متفاوت، نتایج پیچیده‌تر بودند. انتقال کامل و نسخه‌های مقیاس‌شده در ارزیابی اولیه موفقیت صفر درصد داشتند، زیرا برخی مقادیر Q منتقل‌شده عامل را به اعمالی هدایت می‌کردند که در ساختار جدید نامناسب بودند. با ادامه آموزش، این عامل‌ها توانستند رفتار خود را اصلاح کنند و در پایان به موفقیت ۱۰۰ درصد و مسیرهای حدود ۳۲ گام برسند. این نتیجه نشان می‌دهد انتقال منفی لزوماً به شکست نهایی منجر نمی‌شود، اما می‌تواند در شروع آموزش عامل را گمراه کند و هزینه یادگیری اولیه را افزایش دهد. جدول نتایج نیز نشان می‌دهد که در مقصد متفاوت، آموزش از صفر موفقیت اولیه ۳ درصد و موفقیت نهایی ۹۹ درصد داشت، در حالی که انتقال کامل از صفر درصد آغاز شد ولی در پایان به موفقیت ۱۰۰ درصد رسید .

روش انتقال انتخابی با مقایسه همسایگی محلی حالت‌ها، فقط حالت‌هایی را منتقل کرد که ساختار اطراف آن‌ها میان مبدأ و مقصد بدون تغییر تشخیص داده شد. در سناریوی different/selectiveتعداد ۱۸۰ حالت، معادل حدود ۳۱.۹۱ درصد فضای حالت قابل انتقال، کپی شدند و سایر حالت‌های مشکوک حذف شدند. این روش موفقیت اولیه را به ۲۸ درصد رساند که ۲۵ واحد درصد بیشتر از آموزش از صفر بود. بنابراین از نظر عملکرد کاملاً اولیه، انتقال انتخابی شروع ایمن‌تر و مفیدتری ایجاد کرد.

بااین‌حال، تحلیل ۱۰۰ اپیزود نخست نشان داد که همین سناریو نیز نمونه‌ای از انتقال منفی دارد. میانگین پاداش ۱۰۰ اپیزود اول آن برابر 822.05-بود که حدود 40.90واحد کمتر از آموزش از صفر بود. بنابراین، موفقیت اولیه بالاتر به‌تنهایی برای مثبت دانستن انتقال

کافی نیست. ممکن است عامل در تعدادی از ارزیابی‌های اولیه به هدف برسد، اما هنگام ادامه اکتشاف و آموزش، دانش منتقل‌شده باعث برخورد بیشتر با موانع، خانه‌های جریمه یا مسیرهای نامناسب شود. این نتیجه اهمیت تفکیک چهار معیار «عملکرد اولیه»، «سرعت یادگیری»، «عملکرد نهایی» و «انتقال منفی» را نشان می‌دهد.

نقشه تغییر بیشینه مقدار Q پیش و پس از آموزش مقصد نیز نشان داد که عامل مجبور شده است بخش بزرگی از دانش منتقل‌شده را اصلاح کند. در برش بررسی‌شده با has_key=0 و gate_phase=0، عمل حریصانه در ۱۱۰ حالت از ۱۴۱ حالت قابل‌مقایسه، معادل ۷۸.۰۱ درصد، تغییر کرد. این درصد بالا نشان می‌دهد حتی انتقال انتخابی بر پایه شباهت محلی نمی‌تواند تضمین کند که ارزش بلندمدت و بهترین عمل در مقصد بدون تغییر باقی مانده است. ساختار کلی مسیر، محل کلید، خانه‌های جریمه و پیامدهای چندمرحله‌ای می‌توانند با وجود شباهت محلی، ارزش یک حالت را تغییر دهند. در نهایت، عامل انتخابی به موفقیت ۱۰۰ درصد، پاداش ۷۴.۸۵ و میانگین ۳۳.۰۵ گام رسید؛ بنابراین Q-Learning توانست اثر نامناسب دانش اولیه را با تجربه‌های جدید اصلاح کند .

نتیجه کلی انتقال یادگیری این است که انتقال کامل در محیط‌های بسیار مشابه می‌تواند بسیار مؤثر باشد، اما با افزایش تفاوت محیط، خطر انتقال مقادیر Q نامعتبر افزایش می‌یابد. مقیاس‌کردن Q شدت دانش منتقل‌شده را کاهش می‌دهد، ولی مشکل جهت نادرست سیاست را به‌طور کامل حل نمی‌کند. انتقال انتخابی شروع محافظه‌کارانه‌تری فراهم می‌کند، اما معیار شباهت محلی به‌تنهایی برای جلوگیری کامل از انتقال منفی کافی نیست. یک روش قوی‌تر باید علاوه بر ساختار محلی، عدم قطعیتQ ، میزان بازدید حالت، شباهت انتقال‌ها، شباهت پاداش و سازگاری چندمرحله‌ای مسیرها را نیز در تصمیم انتقال دخالت دهد.

11-8. نتیجه‌گیری درباره رابط گرافیکی، آزمون‌ها و بازتولیدپذیری

رابط گرافیکی Pygame نقش مهمی در قابل مشاهده‌کردن رفتار عامل داشت. کاربر می‌تواند محیط، مدل و حالت آموزش یا ارزیابی را انتخاب کند، اجرا را آغاز یا متوقف کند، حرکت را به‌صورت تک‌گام مشاهده کند، سرعت انیمیشن را تغییر دهد و سیاست حریصانه را روی نقشه ببیند. نمایش هم‌زمان شماره اپیزود، epsilon، تعداد گام، پاداش تجمعی، وضعیت کلید، فاز دروازه، عمل انتخاب‌شده، عمل واقعی و خطای TD باعث می‌شود تفاوت میان تصمیم عامل و انتقال تصادفی محیط قابل بررسی باشد. این رابط فقط یک بخش نمایشی نیست؛ بلکه ابزاری برای اشکال‌زدایی، تحلیل مسیر و بررسی حالت‌هایی است که عامل در آن‌ها رفتار غیرمنتظره دارد.

در حالتTraining ، تغییرepsilon ، نرخ موفقیت اخیر و خطای TD امکان مشاهده روند یادگیری را فراهم می‌کنند. در حالت Evaluation، اکتشاف و به‌روزرسانی Q متوقف می‌شود و عامل بر اساس سیاست ذخیره‌شده حرکت می‌کند. تفاوت میان عمل انتخاب‌شده و عمل واقعی نیز اثر مدل انتقال تصادفی 0.8/0.1/0.1را نشان می‌دهد. به این ترتیب، رابط گرافیکی مفاهیم نظری مانند سیاست رفتاری، سیاست حریصانه، خطایTD ، انتقال احتمالی و وابستگی سیاست به کلید و فاز دروازه را به رفتار بصری قابل مشاهده تبدیل کرده است.

مجموعه آزمون‌های واحد نیز بخش‌های اصلی پروژه را پوشش می‌دهد. رفتار محیط، برخورد با دیوار و مرز، جمع‌آوری کلید، در قفل، دروازه دوره‌ای، خانه‌های جریمه، احتمال انتقال، همگرایی Value Iteration ، به‌روزرسانیQ-Learning ، برنامه‌هایepsilon ، ردپای SARSA و حالت‌های مختلف انتقال یادگیری آزموده شده‌اند. آخرین اجرای ثبت‌شده شامل ۲۹ آزمون موفق بود. این موضوع احتمال خطاهای پنهان در بخش‌های اصلی را کاهش می‌دهد و اعتماد به نتایج گزارش‌شده را افزایش می‌دهد. گزارش همچنین دستورهای نصب وابستگی‌ها، اجرای آزمایش‌ها، تولید شکل‌ها، اجرای رابط و اجرای آزمون‌ها را ارائه می‌کند؛ بنابراین نتایج از مسیرهای ذخیره‌شده وSeed های مشخص قابل بازتولید هستند .

11-9. محدودیت‌های نتایج

با وجود کامل‌بودن خط لوله اجرایی، نتایج این پروژه دارای چند محدودیت هستند. نخست، بسیاری از مقایسه‌های اصلی الگوریتم‌ها با Seedپایه ۸ انجام شده‌اند. هرچند مقایسه Sparse و Shaped با سه Seed مستقل گسترش یافت، سه اجرا برای نتیجه‌گیری آماری قوی کافی نیست. نوسان ناشی از اکتشافepsilon-greedy ، انتقال احتمالی محیط و تولید نقشه می‌تواند بر سرعت یادگیری، پاداش و سیاست نهایی اثر بگذارد. بنابراین، بهتر است آزمایش‌های اصلی با حداقل ۱۰ Seed مستقل تکرار شوند و میانگین، انحراف معیار و فاصله اطمینان ۹۵ درصد گزارش شود.

دوم، بودجه آموزش Q-Learning و SARSA برابر ۶۰۰ اپیزود بوده است. موفقیت بالا نشان می‌دهد این بودجه برای یافتن یک سیاست قابل استفاده کافی بوده، اما درصد توافق محدود با Value Iteration و تفاوت سیاست‌های Sparse و Shaped نشان می‌دهد همه مقادیر Q لزوماً به‌طور کامل تثبیت نشده‌اند. افزایش تعداد اپیزودها، کاهش تدریجی‌تر نرخ یادگیری و استفاده از برنامه‌های اکتشاف تطبیقی می‌تواند به تثبیت بیشتر سیاست کمک کند.

سوم، معیار توافق سیاست فقط بهترین عمل انتخاب‌شده را مقایسه می‌کند. در حالت‌هایی که دو یا چند عمل Q تقریباً برابر دارند، تفاوت کوچک عددی باعث اختلاف کامل در معیار argmaxمی‌شود، حتی اگر کیفیت اعمال تقریباً یکسان باشد. در نسخه‌های بعدی بهتر است علاوه بر توافق دقیق، فاصله میان بردارهایQ ، توافق میان مجموعه اعمال تقریباً بهینه و تفاوت ارزش مورد انتظار مسیر نیز بررسی شود.

چهارم، نتایج انتقال انتخابی نشان دادند که شباهت همسایگی محلی برای تشخیص کامل قابلیت انتقال کافی نیست. یک حالت ممکن است در اطراف خود دیوارهای یکسانی داشته باشد، ولی به دلیل تغییر محل کلید، جریمه‌ها یا ساختار دورتر نقشه، ارزش بلندمدت متفاوتی پیدا کند. بنابراین، انتقال مبتنی بر مدل، انتقال دارای وزن اطمینان، انتقال بر اساس شباهت ویژگی‌ها یا روش‌هایی که ابتدا Q منتقل‌شده را با چند اپیزود ارزیابی می‌کنند، می‌توانند انتقال منفی را کاهش دهند.

پنجم، محیط اصلی در اندازه ۱۵×۱۵ بررسی شده است. آزمایش روی نقشه‌های کوچک‌تر و بزرگ‌تر، تراکم‌های مختلف دیوار، دوره‌های متفاوت دروازه و سطوح مختلف تصادفی‌بودن انتقال می‌تواند تعمیم‌پذیری الگوریتم‌ها را بهتر مشخص کند. همچنین اندازه‌گیری مصرف حافظه، تعداد حالت‌های واقعاً بازدیدشده و هزینه محاسباتی به‌ازای هر اپیزود، مقایسه کامل‌تری از مقیاس‌پذیری روش‌ها فراهم می‌کند.

11-10. پیشنهاد برای ادامه کار

در ادامه این پروژه، نخست پیشنهاد می‌شود تمام تنظیم‌های اصلی باSeed های بیشتر اجرا و نتایج با فاصله اطمینان ۹۵ درصد گزارش شوند. دوم، می‌توان نرخ یادگیری α را به‌صورت کاهشی یا وابسته به تعداد بازدید حالت–عمل تنظیم کرد تاQ ها در مراحل پایانی آموزش نوسان کمتری داشته باشند. سوم، استفاده از روش‌هایی مانند Double Q-Learning می‌تواند بیش‌برآورد ناشی از عمل maxرا کاهش دهد. چهارم، برای SARSA می‌توان replacing trace را با accumulating trace و true online SARSA(λ)مقایسه کرد تا اثر نوع ردپا بر پایداری و سرعت یادگیری مشخص شود.

برای انتقال یادگیری نیز می‌توان برای هر حالت یک امتیاز اطمینان تعریف کرد و Q منتقل‌شده را متناسب با شباهت ساختاری، تعداد بازدید در مبدأ و اختلاف پاداش‌ها وزن‌دهی کرد. همچنین عامل می‌تواند در چند اپیزود اولیه، میان Q منتقل‌شده و Q صفر ترکیب تطبیقی انجام دهد و هرجا شواهد مقصد با دانش مبدأ ناسازگار بود، وزن انتقال را کاهش دهد. ثبت نمودار زمان‌مند تغییرQ ، تعداد حالت‌های اصلاح‌شده و مدت لازم برای جبران انتقال منفی نیز تحلیل دقیق‌تری از سازگاری عامل فراهم می‌کند.

در بخش محیط، می‌توان دوره دروازه را متغیر کرد، بیش از یک کلید یا در قفل افزود، برخی موانع را متحرک ساخت و چند هدف با پاداش متفاوت تعریف کرد. این تغییرها فضای حالت را بزرگ‌تر می‌کنند و امکان بررسی روش‌های تقریب تابع، شبکه عصبی و الگوریتم‌های Deep Reinforcement Learning را فراهم می‌سازند. با بزرگ‌شدن فضای حالت، ذخیره جدول Q برای تمام ترکیب‌ها پرهزینه خواهد شد و استفاده از DQN یا روش‌های actor–critic می‌تواند مسیر طبیعی توسعه پروژه باشد.

11-11. نتیجه‌گیری نهایی

نتیجه اصلی این پروژه آن است که هیچ الگوریتم یا تنظیمی به‌تنهایی در تمام معیارها بهترین نیست Value Iteration .با دسترسی به مدل کامل، یک مرجع دقیق و قابل تفسیر برای تحلیل ارزش و سیاست فراهم کرد، اما هزینه آن با افزایش افق تنزیل و اندازه فضای حالت بیشتر می‌شود Q-Learning .بدون نیاز به مدل توانست به موفقیت کامل برسد و برنامه کاهش خطی epsilon نسبت به کاهش نمایی نتیجه متعادل‌تری ایجاد کرد SARSA(λ) .نشان داد که انتقال خطای TD به تصمیم‌های گذشته می‌تواند یادگیری را بهبود دهد و λ برابر ۰.۳ در این مسئله بهترین توازن میان موفقیت، پاداش، طول مسیر و زمان آموزش را داشت. پاداش Shaped عموماً سرعت یادگیری را افزایش داد، ولی اثر آن بر کیفیت مسیر به الگوریتم وابسته بود. انتقال یادگیری در محیط مشابه بسیار سودمند بود، اما در محیط متفاوت خطر انتقال منفی را آشکار کرد و نشان داد که شباهت ظاهری یا محلی به‌تنهایی تضمین‌کننده مفیدبودن دانش مبدأ نیست.

در مجموع، پروژه توانست چرخه کامل تعریفMDP ، طراحی محیط، پیاده‌سازی الگوریتم، تنظیم پارامتر، آموزش، ارزیابی، تحلیل سیاست، انتقال دانش، نمایش گرافیکی، آزمون نرم‌افزار و بازتولید نتایج را پوشش دهد. مهم‌ترین برداشت عملی این است که ارزیابی یک عامل نباید فقط بر نرخ موفقیت نهایی متکی باشد. سرعت رسیدن به موفقیت، پاداش، طول مسیر، پایداری میان اجراها، میزان تغییر سیاست، رفتار اولیه پس از انتقال و هزینه محاسباتی نیز باید هم‌زمان بررسی شوند. همین نگاه چندمعیاره باعث شد تفاوت‌های مهمی میان تنظیم‌هایی آشکار شود که در ظاهر همگی به موفقیت نزدیک به ۱۰۰ درصد رسیده بودند. بنابراین، ارزش اصلی این پروژه علاوه بر حل هزارتو، ارائه یک چارچوب تحلیلی برای فهم علت موفقیت، شکست، نوسان و سازگاری الگوریتم‌های یادگیری تقویتی در یک محیط پویا و تصادفی است.

مخزن عمومی پروژه

https://github.com/mina2collab/RL_FinalProject_40403384

۱۲. استفاده از ابزار هوش مصنوعی

از ابزار هوش مصنوعی به‌عنوان دستیار ساختاردهی اولیه، اشکال‌زدایی، پیشنهاد آزمون و نگارش مستندات استفاده شد. همه پیشنهادها توسط دانشجو بررسی، اجرا و با خروجی واقعی اعتبارسنجی شدند. هیچ نتیجه عددی بدون اجرای واقعی کد وارد گزارش نشده است و نمودارهای گزارش نیز از داده‌های خام تولیدشده توسط پروژه ساخته شده‌اند