به نام او

درس یادگیری تقویتی

پروژه دوم، پوریا صفائی ۹۸۱۱۰۴۰۲

تابع $valueiteration$ و $policyiteration$ در کلاس $FrozenLake$ وظیفه به دست آوردن سیاست بهینه و مقدار تابع ارزش حالت ها را برای هر بخش سوال دارد. ابتدا یک شی از این کلاس با توجه به جدول داده شده در سوال ساخته و به بخش‌های مختلف پاسخ میدهیم. در اینجا در ر دو الگوریتم مفدار گاما برابر با ۹.۰ فرض شده و مقدار اپسیلون(برای بررسی همگرایی) عدد $۱۰^{-۶}$ فرض شده. مقدار تعداد گام های هر الگوریتم(ایتریشن) در کنار پاسخ هر بخش ارائه شده که بیانگر تعداد تکرار حلقه تا همگرایی تابع ارزش عمل/سیاست میباشد. .(تمام کدهای مربوط در فایل $project۲.py$ در فایل زیپ قرار داده شده است.)

سوال ۱

همانطور که میتوان مشاهده کرد، طبق الگوریتم $valueiteration$، مقادیر تابع ارزش برای تمام استیت های جدول محاسبه شده است. از طرفی سیاست بهینه نیز برای تمام استیت ها به دست آمده. تعداد کل پیمایش ها(جمع تکرار لوپ روی تابع ارزش تا همگرایی آن و سپس همگرایی سیاست) برابر ۷۹۲ میباشد و این الگوریتم ۵ بار اجرا شده تا سیاست نیز همگرا شود.

```
Python 3.10.2 (tags/v3.10.2:a58ebcc, Jan 17 2022, 14:12:
Value Iteration Algorithm Policy Number of Iterations:
792

Optimal Value Iteration Algorithm Policy:
[[1 1 1 0]
 [1 1 1 0]
 [1 2 1 0]
 [1 0 1 1]
 [2 2 2 0]]

Optimal Value Iteration Algorithm Value Function:
[[ 5.31440179  5.90489179  6.56099179  5.90489261]
 [ 5.90489179  6.56099179  7.28999179 -9.99999179]
 [ 6.56099179  7.28999179  8.09999179 -9.99999179]
 [ 7.28999179 -9.99999179  8.99999179  9.99999179]
 [ 8.09999179  8.99999179  9.99999179  9.99999179]]
```

سوال ۲

همانطور که میتوان مشاهده کرد، طبق الگوریتم $policyiteration$، مقادیر تابع ارزش برای تمام استیت های جدول محاسبه شده است. از طرفی سیاست بهینه نیز برای تمام استیت ها به دست آمده. تعداد کل پیمایش ها(جمع تکرار لوپ روی تابع ارزش تا همگرایی آن و سپس همگرایی سیاست) برابر ۱۳۲ میباشد و همانطور که میتوان مشاهده کرد، تعداد پیمایش های این الگوریتم کمتر از الگوریتم قبلی است. با مقایسه سیاست بهینه و مقادیر تابع ارزش بهینه با بخش قبل، میتوان مشاهده کرد که این مقادیر برای هر دو بخش با یکدیگر یکسان میباشد.

```
Policy Iteration Algorithm Policy Number of Iterations:
132

Optimal Policy Iteration Algorithm Policy:
[[1 1 1 0]
 [1 1 1 0]
 [1 2 1 0]
 [1 0 1 1]
 [2 2 2 0]]

Optimal Policy Iteration Algorithm Value Function:
[[ 5.31440179  5.90489179  6.56099179  5.90489261]
 [ 5.90489179  6.56099179  7.28999179 -9.99999179]
 [ 6.56099179  7.28999179  8.09999179 -9.99999179]
 [ 7.28999179 -9.99999179  8.99999179  9.99999179]
 [ 8.09999179  8.99999179  9.99999179  9.99999179]]
```