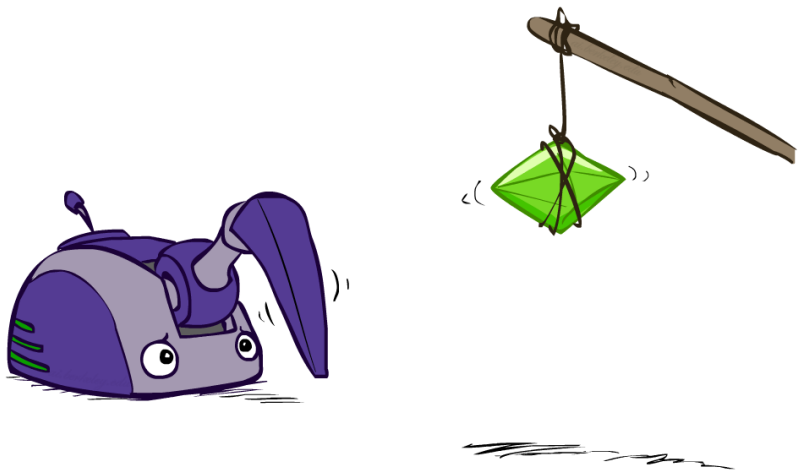
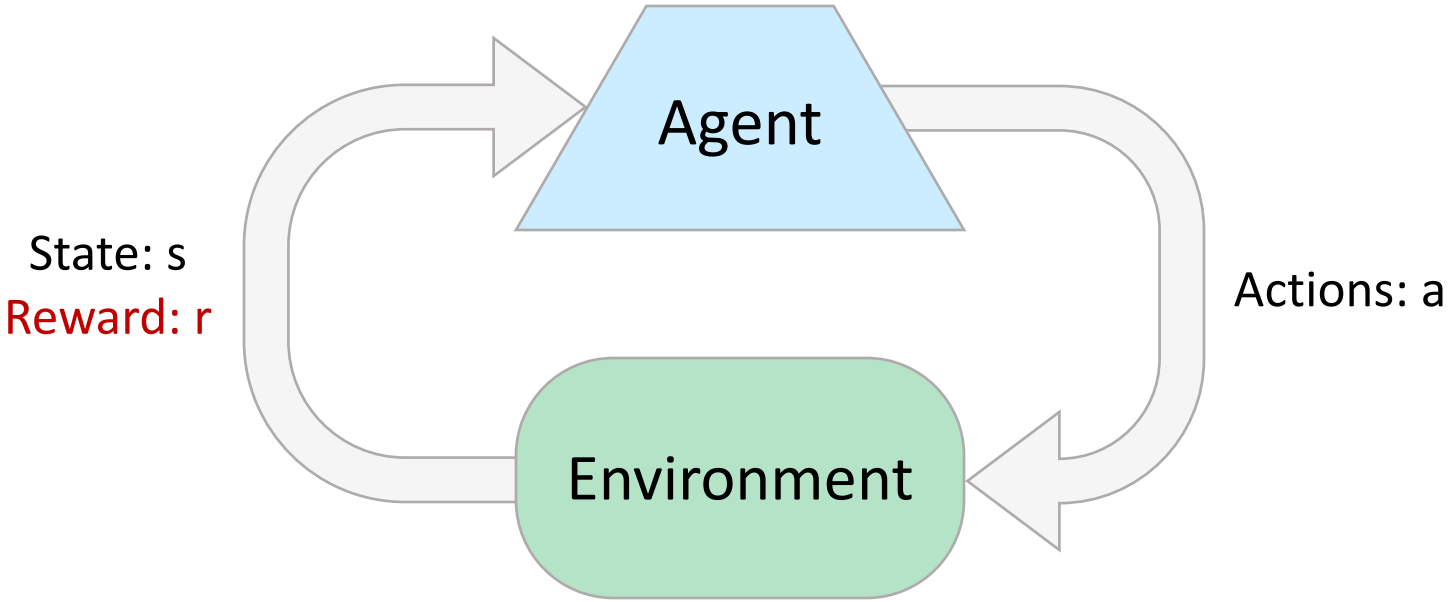
یادگیری تقویتی
(Reinforcement Learning)

1. فرایند تصمیم مارکوف (Markov Decision Processes)
2. الگوریتم تکرار مقدار (Value Iteration)
3. الگوریتم تکرار سیاست (Policy Iteration)
4. یادگیری تقویتی (Reinforcement Learning)

یادگیری تقویتی
(Reinforcement Learning)



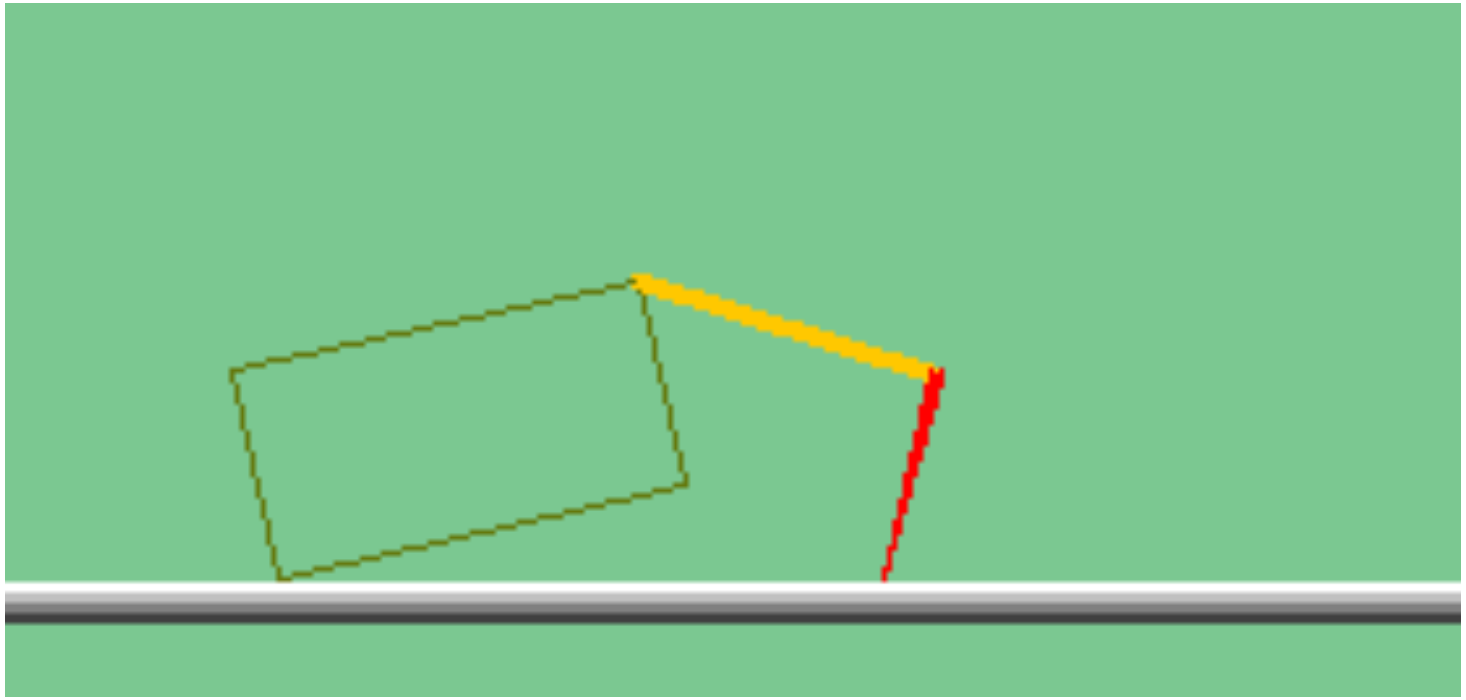
یادگیری تقویتی



❑ایده‌ی اصلی.

- دریافت بازخورد از محیط به شکل پاداش‌ها.
- سودمندی عامل به وسیله‌ی تابع پاداش تعریف می‌شود.
- عامل باید به گونه‌ای عمل کند که سودمندی مورد انتظار را بیشینه سازد.
- یادگیری بر مبنای نمونه‌های مشاهده شده از پیامدها است!

روبات خزنده!



[Demo: Crawler Bot (L10D1)]

یادگیری تقویتی -- مرور کلی

❑ یادگیری تقویتی منفعل (Passive RL) :

یاد گرفتن از تجربه‌ها (عامل با یک سیاست از پیش‌مشخص‌شده حرکت می‌کند و هدف آن ارزیابی ارزش حالت‌ها است، نه تصمیم‌گیری)

1. یادگیری مبتنی بر مدل :

1. عامل ابتدا مدل MDP (یعنی Tو R) را بر پایه‌ی تجربه‌های خود تخمین می‌زند. سپس با استفاده از این مدل تقریبی، ارزش حالت‌ها را محاسبه می‌کند.

2. یادگیری مستقل از مدل:

عامل فقط بر پایه‌ی تجربه، بدون دانستن مدل، ارزش حالت‌ها را محاسبه می‌کند.(مستقیماً ارزش هر حالت را یاد می‌گیرد)

Value Learning

1. **ارزیابی مستقیم:** محاسبه‌ی میانگین مجموع پاداش‌ها در اپیزودها
2. **یادگیری تفاضل زمانی:** به‌روزرسانی تدریجی ارزش هر حالت با استفاده از ارزش حالت بعدی

❑ یادگیری تقویتی فعال (Active RL):

عامل خودش باید تصمیم بگیرد و تجربه جمع‌آوری کند (عامل دیگر فقط یک سیاست را دنبال نمی‌کند، بلکه باید خودش سیاست بهینه را کشف کند. بنابراین، علاوه بر یادگیری ارزش‌ها، باید انتخاب عمل نیز انجام دهد.)

- **یادگیری مستقل از مدل :**

Q-Learning

- **Q-Learning:** به‌روزرسانی تدریجی ارزش هر Q-state با استفاده از بیشینه‌ی ارزش Q-state های حالت بعدی
- **Approximate Q-Learning:** به‌روزرسانی تدریجی ارزش هر Q-state با استفاده از یک تابع تقریب

- **چالش‌های اصلی:**
  - چگونه به‌صورت مؤثر محیط را کاوش کند؟
  - چگونه بین کاوش (exploration) و بهره‌برداری (exploitation) تعادل برقرار کند؟

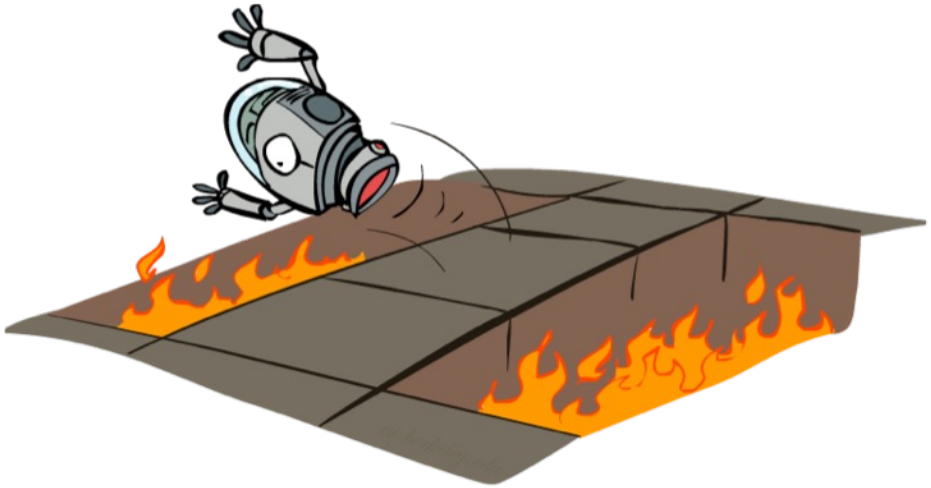
یادگیری تقویتی

- فرض می‌کنیم یک فرایند تصمیم مارکوف (MDP) داریم:
  - یک مجموعه از حالت‌ها $s \in S$
  - یک مجموعه از اعمال A
  - یک مدل $T(s,a,s')$
  - یک تابع پاداش $R(s,a,s')$

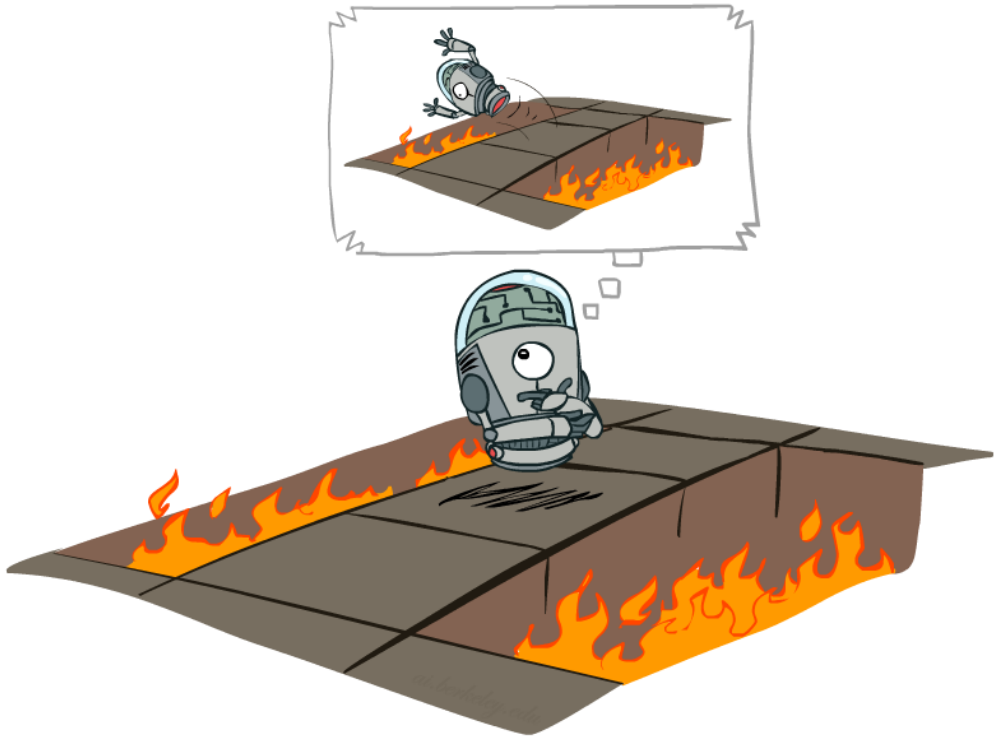
- همچنان به دنبال یک سیاست $\pi(s)$ هستیم

- تفاوت: توابع T و R ناشناخته هستند.
  - یعنی، نمی‌دانیم کدام حالت‌ها بهتر هستند و نتیجه‌ی هر عمل چیست؟
  - برای یادگیری باید عمل‌های مختلف و حالت‌های نتیجه شده را آزمایش کنیم.

آفلاین (MDP) یا آنلاین (یادگیری تقویتی)؟



یادگیری انلاین



راه حل آفلاین

یادگیری تقویتی -- مرور کلی

❑ یادگیری تقویتی منفعل (Passive RL) :

یاد گرفتن از تجربه‌ها (عامل با یک سیاست از پیش‌مشخص‌شده حرکت می‌کند و هدف آن ارزیابی ارزش حالت‌ها است، نه تصمیم‌گیری)

1. یادگیری مبتنی بر مدل :

1. عامل ابتدا مدل MDP (یعنی T و R) را بر پایه‌ی تجربه‌های خود تخمین می‌زند. سپس با استفاده از این مدل تقریبی، ارزش حالت‌ها را محاسبه می‌کند.

2. یادگیری مستقل از مدل:

عامل فقط بر پایه‌ی تجربه، بدون دانستن مدل، ارزش حالت‌ها را محاسبه می‌کند.(مستقیماً ارزش هر حالت را یاد می‌گیرد)

Value Learning

1. **ارزیابی مستقیم:** محاسبه‌ی میانگین مجموع پاداش‌ها در اپیزودها
2. **یادگیری تفاضل زمانی:** به‌روزرسانی تدریجی ارزش هر حالت با استفاده از ارزش حالت بعدی

❑ یادگیری تقویتی فعال (Active RL):

عامل خودش باید تصمیم بگیرد و تجربه جمع‌آوری کند (عامل دیگر فقط یک سیاست را دنبال نمی‌کند، بلکه باید خودش سیاست بهینه را کشف کند. بنابراین، علاوه بر یادگیری ارزش‌ها، باید انتخاب عمل نیز انجام دهد.)

- **یادگیری مستقل از مدل :**

Q-Learning

- **Q-Learning:** به‌روزرسانی تدریجی ارزش هر Q-state با استفاده از بیشینه‌ی ارزش Q-state های حالت بعدی
- **Approximate Q-Learning:** به‌روزرسانی تدریجی ارزش هر Q-state با استفاده از یک تابع تقریب

- **چالش‌های اصلی:**
  - چگونه به‌صورت مؤثر محیط را کاوش کند؟
  - چگونه بین کاوش (exploration) و بهره‌برداری (exploitation) تعادل برقرار کند؟

یادگیری تقویتی منفعل - مبتنی بر مدل

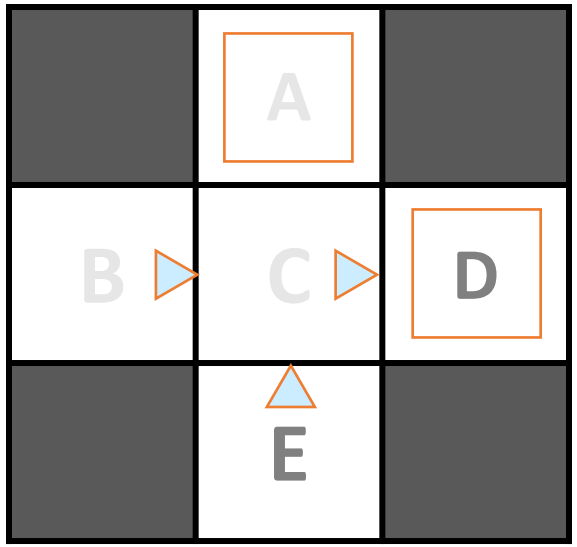
- ایده‌ی یادگیری مبتنی بر مدل:
  - یادگیری یک مدل تقریبی بر مبنای تجربیات
  - به دست‌آوردن مقادیر طوریکه انگار مدل یادگرفته‌شده صحیح است

- گام اول: یادگیری یک مدل تجربی از MDP
  - شمارش تعداد s’ ها به ازای هر s, a
  - نرمال‌سازی برای به دست آوردن مقدار تقریبی $\hat{T}(s,a,s')$
  - یادگیری مقادیر $\hat{R}(s,a,s')$ با توجه به تجربه‌های (s, a, s')

- گام دوم: حل کردن MDP یاد گرفته شده.
  - مثلا با استفاده از الگوریتم تکرار مقدار.

مثال: یادگیری مبتنی بر مدل

سیاست ورودی π



Assume: $\gamma = 1$

اپیزودهای مشاهده شده (آموزش)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 3

E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4

E, north, C, -1
C, east, A, -1
A, exit, x, -10

مدل یادگرفته شده

$\hat{T}(s, a, s')$

T(B, east, C) = 1.00
T(C, east, D) = 0.75
T(C, east, A) = 0.25
...

$\hat{R}(s, a, s')$

R(B, east, C) = -1
R(C, east, D) = -1
R(D, exit, x) = +10
...

یادگیری تقویتی -- مرور کلی

❑ یادگیری تقویتی منفعل (Passive RL) :

یاد گرفتن از تجربه‌ها (عامل با یک سیاست از پیش‌مشخص‌شده حرکت می‌کند و هدف آن ارزیابی ارزش حالت‌ها است، نه تصمیم‌گیری)

1. **یادگیری مبتنی بر مدل :**
   1. عامل ابتدا مدل MDP (یعنی T و R) را بر پایه‌ی تجربه‌های خود تخمین می‌زند. سپس با استفاده از این مدل تقریبی، ارزش حالت‌ها را محاسبه می‌کند.
2. **یادگیری مستقل از مدل:**

عامل فقط بر پایه‌ی تجربه، بدون دانستن مدل، ارزش حالت‌ها را محاسبه می‌کند.(مستقیماً ارزش هر حالت را یاد می‌گیرد)

Value Learning

1. **ارزیابی مستقیم:** محاسبه‌ی میانگین مجموع پاداش‌ها در اپیزودها
2. **یادگیری تفاضل زمانی:** به‌روزرسانی تدریجی ارزش هر حالت با استفاده از ارزش حالت بعدی

❑ یادگیری تقویتی فعال (Active RL):

عامل خودش باید تصمیم بگیرد و تجربه جمع‌آوری کند (عامل دیگر فقط یک سیاست را دنبال نمی‌کند، بلکه باید خودش سیاست بهینه را کشف کند. بنابراین، علاوه بر یادگیری ارزش‌ها، باید انتخاب عمل نیز انجام دهد.)

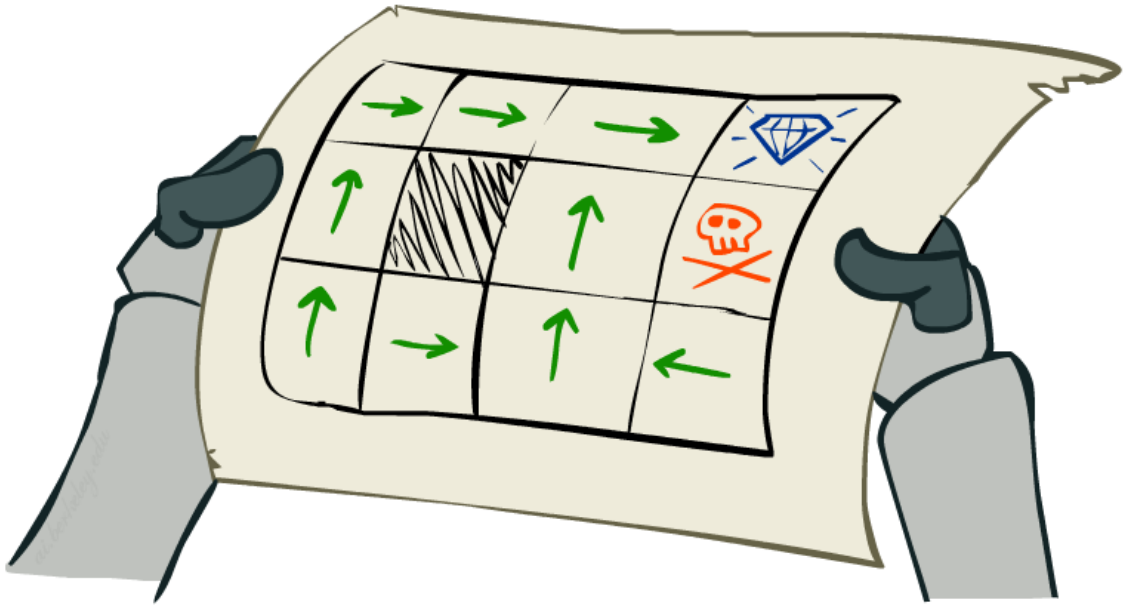
- **یادگیری مستقل از مدل :**

Q-Learning

- **Q-Learning:** به‌روزرسانی تدریجی ارزش هر Q-state با استفاده از بیشینه‌ی ارزش Q-state های حالت بعدی
- **Approximate Q-Learning:** به‌روزرسانی تدریجی ارزش هر Q-state با استفاده از یک تابع تقریب

- **چالش‌های اصلی:**
  - چگونه به‌صورت مؤثر محیط را کاوش کند؟
  - چگونه بین کاوش (exploration) و بهره‌برداری (exploitation) تعادل برقرار کند؟

یادگیری تقویتی منفعل - مستقل از مدل



- وظیفه‌ی ساده شده: ارزیابی سیاست!
  - ورودی: یک سیاست ثابت $\pi(s)$
  - تابع تغییر حالت $T(s,a,s')$ ناشناخته است
  - تابع پاداش $R(s,a,s')$ ناشناخته است.
  - هدف: یادگیری ارزش حالت‌ها

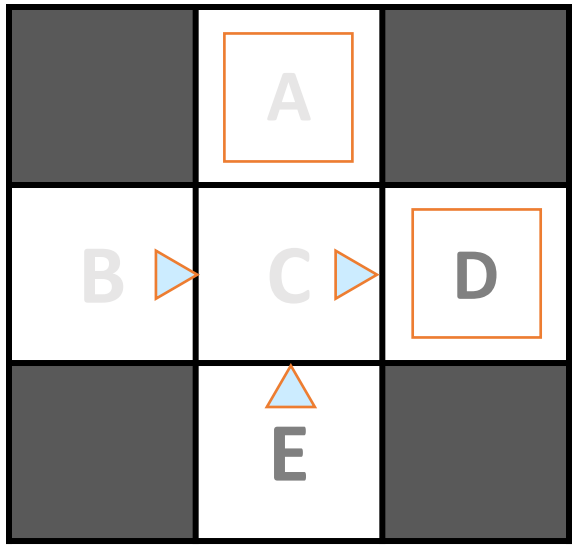
- در این مورد:
  - عامل خودش تصمیم‌گیر نیست، فقط همراه مسیر داده‌شده حرکت می‌کند.
  - هیچ انتخابی برای انجام اعمال ندارد.
    - فقط باید سیاست را اجرا کند و از تجربه‌ها یاد بگیرد
    - این برنامه‌ریزی آفلاین نیست! زیرا عامل واقعا در محیط عمل می‌کند.

ارزیابی مستقیم

- **هدف:** محاسبه‌ی ارزش هر حالت تحت سیاست ثابت π
- **ایده:** میانگین‌گیری از مقادیر نمونه‌ی مشاهده شده.
  - بر طبق π عمل کن
  - هر بار که با یک حالت روبرو می‌شوی، محاسبه کن که مجموع (کاهش یافته) پاداش‌ها چقدر باید باشند.
  - از نمونه‌های مشاهده شده میانگین بگیر.
- این روش ارزیابی مستقیم (direct evaluation) نام دارد

مثال: ارزیابی مستقیم

سیاست ورودی π



Assume: $\gamma = 1$

اپیزودهای مشاهده شده (آموزش)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +10

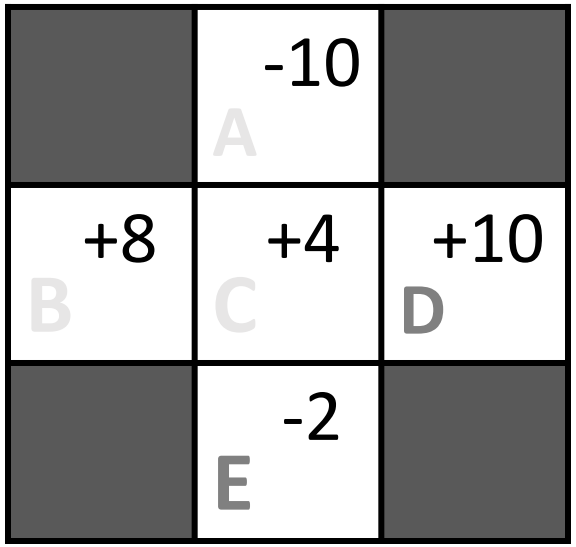
Episode 3

E, north, C, -1
C, east, D, -1
D, exit, x, +10

Episode 4

E, north, C, -1
C, east, A, -1
A, exit, x, -10

مقادیر خروجی



مزایا و معایب ارزیابی مستقیم

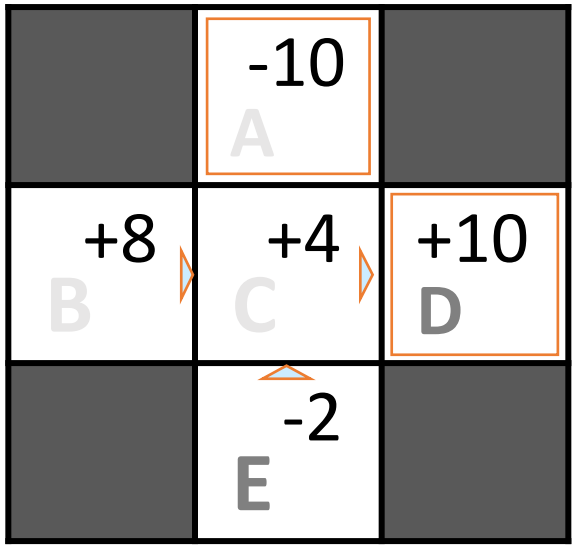
❑مزایا:

- درک ان ساده است.
- نیاز به دانش در مورد T یا R ندارد.
- در نهایت، فقط با استفاده از نمونه تجربه‌ها، میانگین درست ارزش‌ها را محاسبه می‌کند.

❑معایب:

- اتلاف اطلاعات به دلیل در نظر نگرفتن ارتباط میان حالت‌ها.
- یادگیری هر حالت به صورت جداگانه.
- و در نتیجه، نیاز به زمان زیاد برای یادگیری.

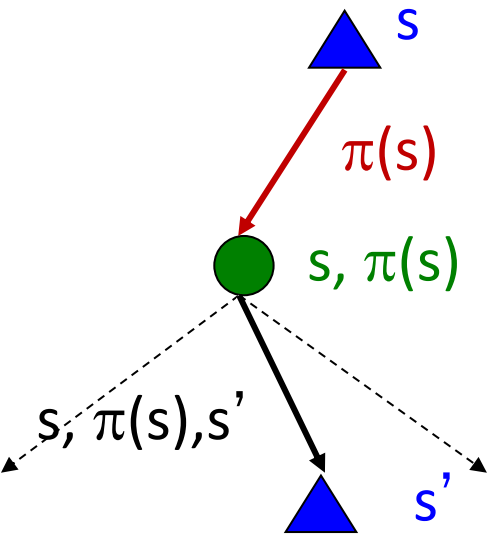
مقادیر خروجی



اگر طبق این سیاست، هم B و هم E به C می‌رسند، چطور ممکن است ارزش آن‌ها با هم متفاوت باشد؟

چرا از روش ارزیابی سیاست استفاده نکنیم؟

- شکل ساده شده‌ی معادله بلمن، ارزش V را برای یک سیاست خاص محاسبه می‌کند.



$$V_0^\pi(s) = 0$$

$$V_{k+1}^\pi(s) \leftarrow \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V_k^\pi(s')]$$

  - این روش، به‌طور کامل ارتباط بین حالت‌ها را در نظر می‌گیرد.
  - ولی متأسفانه، برای انجام این کار به تابع‌های T و R نیاز داریم!

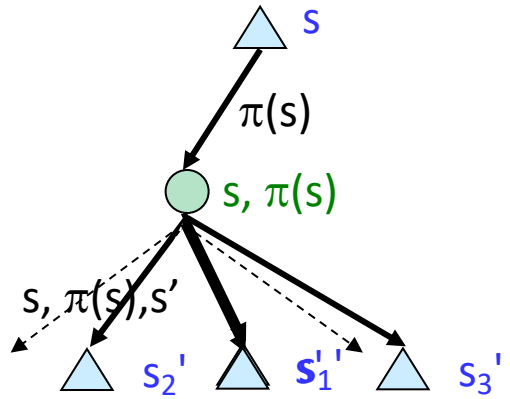
- سوال کلیدی: چگونه می‌توان ارزش حالت‌ها را بدون نیاز به دانستن T و R محاسبه نمود؟
  - در واقع، چطور می‌توانیم میانگین‌گیری وزن‌دار انجام دهیم، بدون اینکه وزن‌ها (یعنی احتمال‌های انتقال) را بدانیم؟

ارزیابی سیاست مبتنی بر نمونه‌برداری؟

- **هدف:** می‌خواهیم تخمین خود را از V با استفاده از معادله‌ی زیر بهبود دهیم:

$$V_{k+1}^{\pi}(s) \leftarrow \sum_{s'} T(s, \pi(s), s')[R(s, \pi(s), s') + \gamma V_k^{\pi}(s')]$$

- **ایده:** از حالت‌های نتیجه s’ نمونه‌برداری کن (با انجام عمل!) و سپس از آن‌ها میانگین بگیر.



نمی‌توانیم زمان را به عقب برگردانیم تا بارها و بارها از حالت s نمونه‌برداری کنیم.

$$sample_1 = R(s, \pi(s), s'_1) + \gamma V_k^{\pi}(s'_1)$$
$$sample_2 = R(s, \pi(s), s'_2) + \gamma V_k^{\pi}(s'_2)$$
$$\dots$$
$$sample_n = R(s, \pi(s), s'_n) + \gamma V_k^{\pi}(s'_n)$$

$$V_{k+1}^{\pi}(s) \leftarrow \frac{1}{n} \sum_i sample_i$$