به نام خدا

الف:

Exploration: در محدوده چیزی که تا به الان فهمیده است به صورت جزیی جستجو میکند. یعنی از دانش فعلی که به دست آورده است بهره میبرد.

Exploitation: با تغییر محدوده به صورت گسترده تر جستجو میکند. یعنی به اکتشاف اعمال جدید میپردازد.

عاملی که فقط Exploit یا فقط Explore میکند موفق عمل نمیکند.

اگر فقط Exploit کند، چون دانش فعلی ما کامل نیست و ممکن است برداشتی که از محیط داریم برداشت دقیقی نباشد پس باید به عامل اجازه اکتشاف و جستجوی اکشن های دیگر را بدهیم تا تجاربش افزایش پیدا کند و دانشش به روز شود.

اگر فقط Explore کند، رفتار آن تصادفی میشود و دانش پیشینی که به دست آورده است استفاده نمیکند پس باید به عامل اجازه استفاده از تجارب گذشته را هم بدهیم.

برای به دست اوردن پاداش زیاد عامل باید یک تعادلی بین Exploitation و Exploration ایجاد کند یعنی نه کاملا تصادفی عمل کند نه کاملا براساس دانش فعلی و چیزی بین این دو مورد باشد.

ب:

ارزش وضعیت: پیش بینی پاداش **تجمعی** یا همان expected discounted sum of rewards میباشد. پس تجمیعی از پاداش ها داریم و تفاوتش با پاداش همین است.

پ:

روش مبتنی بر policy: سیاست بهینه از طریق آموزش مستقیم سیاست به دست می آید. در اینجا نگاشت وضعیت به بهترین عمل یا احتمال مجموعه ای از اعمال ممکن یاد گرفته میشود.

روش مبتنی بر value: به دنبال تابع ارزش بهینه میباشد که منجر به سیاست بهینه میشود. در اینجا نگاشت وضعیت به بهترین ارزش یاد گرفته میشود.

Subject:
Year : Month : Day : ( ) page : ( )

$$R(t) = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots$$

$$R(t) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

$\gamma = 1$

step 0 :

$$R(0) = r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \gamma^4 r_5$$

$$= 2 + 0 - 1 + 2 + 8 = 11$$

step 1 :

$$R(1) = r_2 + \gamma r_3 + \gamma^2 r_4 + \gamma^3 r_5$$

$$= 0 - 1 + 2 + 8 = 9$$

step 2 :

$$R(2) = -1 + 2 + 8 = 9$$

step 3 :

$$R(3) = 2 + 8 = 10$$

Soroush

step 4 :

$R(4) = 8$

$\delta = 0.5$

step 0 :

$$R(0) = 2 + \frac{1}{2} \times 0 + \frac{1}{4} \times -7 + \frac{1}{8} \times 2 + \frac{1}{16} \times 8$$

$$= \frac{5}{2}$$

step 1 :

$$R(1) = 0 + \frac{1}{2} \times -1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 8 = 1$$

step 2 :

$$R(2) = -1 + \frac{1}{2} \times 2 + \frac{1}{4} \times 8 = 2$$

step 3 :

$$R(3) = 2 + \frac{1}{2} \times 8 = 6$$

step 4 :

$R(4) = 8$