تمرین سری 7.2

امیرحسین باقری 9810
5621

سوال اول:

ابتدا فرض کنید که بازی گفته TA تعداد iteration ها بی‌نهایت است.

و همین تعداد iteration ها برای Value تعداد حالات V^*

است، همین تعداد iter ها برای Policy تعداد حالات

در آپدیت های $\pi(s)$ است.

در حالت کلی Policy تعداد iter های کمتری برای converge نیاز دارد

اما هر Policy iter هزینه محاسباتی بیشتری نسبت به Value iteration

دارد.

بنابراین با تعداد iteration های بی‌نهایت روش Value سریع‌تر است.

اثبات:

Value iteration:

for i ≤ iter & ($V^s_{K+1} \neq V^s_K$ for all state) :

$$V_{K+1}(s) = \max_{a} \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^{*}(s')\right]$$

$O(S^2A)$

order of value iteration $\longrightarrow O(nS^2A)$

Policy iteration:

① initialize random $\pi(s)$ first

② iterate until values converge

$$V^{\pi_i}_{K+1}(s) \leftarrow \sum_{s'} T(s,\pi_i(s),s')\left[R(s,\pi_i(s),s') + \gamma V^{\pi_i}_{K}(s')\right]$$

③ extraction until convergence

$$\pi_{i+1}(s) = \arg\max_{a} \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^{\pi_i}(s')\right]$$

optimal, iteration is order of 10

۱۰.

state ③ → $|S| \times |S| \times |a| = (S^2 a)$

state ② → $|S|^3$ order آن در الگوریتم

است.

convergence of $V_{k+1}^{\pi_i}(s) \leftarrow \sum_{s'} P[R + \gamma V_k^{\pi_i}(s')]$

$O(S^3)$ اردر عبارت بالا CSE 473 luke Zettlemoyer ← بر اساس اسلایدهای

است. ← iteration در $O(|S|^3 + |S|^2 a)$ ←

نباید این اگر $|S| > a$ باشد که در اکثریت مواقع هستند.

اردر policy iteration از value iteration در تعداد iter ها کمتر است. بیشتر است.

Policy → $O(N(|S|^3 + |S|^2|A|))$

Value ← $O(N|S|^2|A|)$

⊗ $O(\text{Policy}) > O(\text{iter})$

Value

⊗ در تعداد iter های پایین

Value iteration سریع تر است

در صفحه قبل عنوان نموده‌ایم که |S| از |A| بیشتر باشد آنگاه هم بنا بر اردر نامگذاری است. اما باز هم value سریع‌تر است.

سوال 2

برای حل سوال ⊗ ثابت می‌کنیم که حداقل V و Q را داریم.

Calculating Policy from V

$\pi^*(s) = \arg\max_a Q^*(s,a) \rightarrow |S||A|$

$\hookrightarrow O(|S||A|^a)$

بنابراین استفاده از Q ها بهتر است.

دقت کنید که عملی که دادیم که حداقلی Q و V را داریم.

چون برای Extract Policy باید الگوریتم converge را

در هر iter پیدا کنیم. با یک کار اضافه می‌توان Q را این

پیدا کرد. بنابراین حاسبه Q ها و V ها

از نظر زمانی تفاوت چندانی ندارند.

$O(Q) = O(|S||A|)$

$O(V) = O(|S|^2|A|)$

بنابراین Q بهتر است.