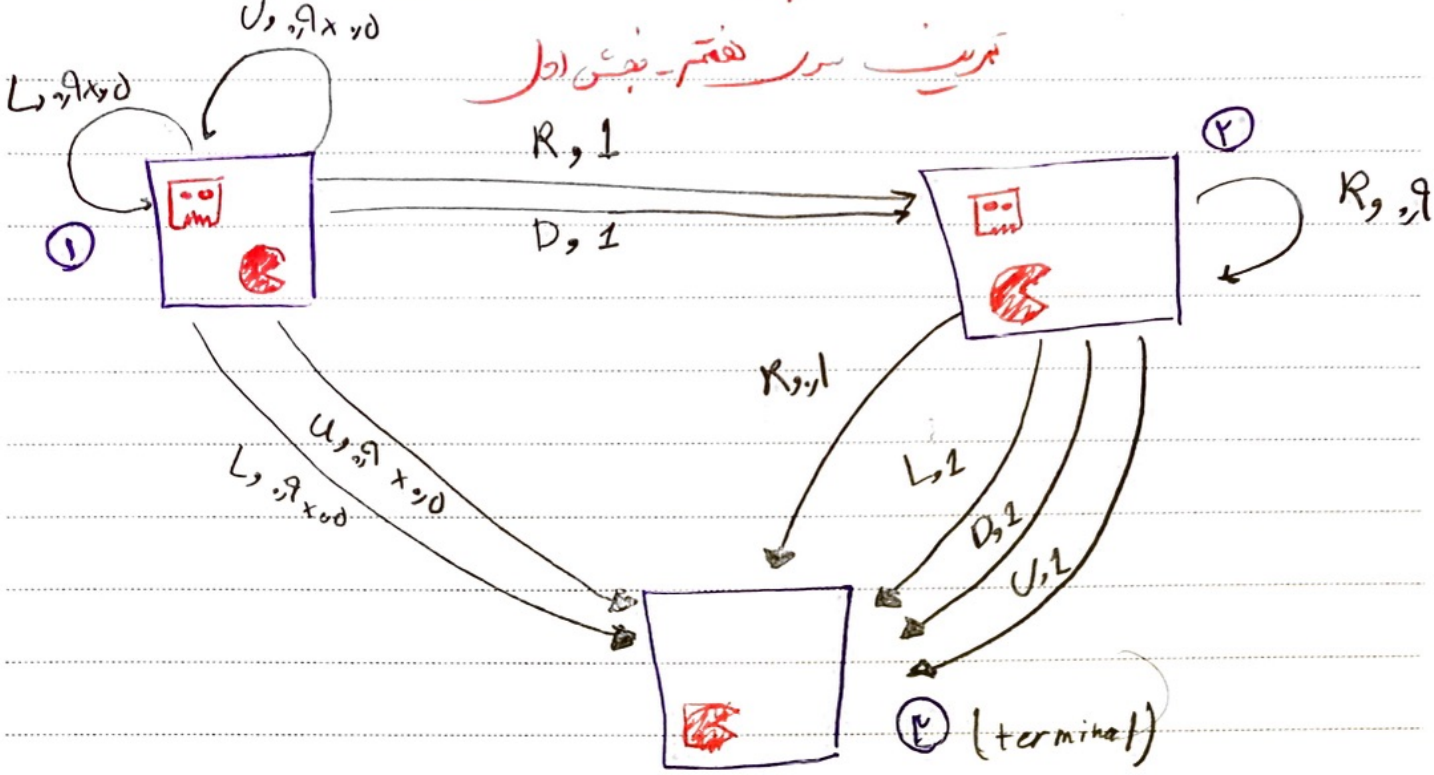
Subject: ۹۸۱۰۳۸۹۷ محمدابول نژادیان

Date

به نام خدا

تمرین سری هفتم - بخش اول



* توجه ① : در حالت ۱، روح با احتمال ۰.۵ به راست و با احتمال ۰.۵ به پایین حرکت می‌کند، زیرا در هر دو حالت به یک متر یک خانه نزدیک‌تر می‌شود.

* توجه ② : تفاوت حالت‌ها با حالاتی که رسم نشده‌اند، این است که فاصله روح تا یک متر از پیش تعیین شده است.

۲) Optimal Policy به شکل زیر تعریف می‌شود :

$\forall s \in S : V^{\pi^*}(s) \geq V^{\pi}(s)$ (π^*: Optimal Policy)

سپس باید بگوییم که معادله bellman-Opt. ① جواب دارد

② یکی از این جواب‌ها از باقی جواب‌ها بزرگتر است

① معادله bellman-Optimization جواب دارد: طبق حرفی که در کلاس راجع به

Subject:
Date

Fixed Point زده شد، این را، حل هاشد روش نیوتن را اصول حتماً به جواب همگرا می شود.

(۲) یکی از جواب ها از باقی جواب ها، بزرگتر است؛

طبق چیزی که در Policy iteration دیدیم، داریم:

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

که چون در هر مرحله داریم argmax می گیریم، قطعاً $V^{\pi_{i+1}} \geq V^{\pi_i}$ به ازای همهٔ s ها.

← حال به دلیل اینکه $a \in A$ و داشته A ($|A|$) محدود است، این بهبود در بدترین حالت، از بدترین π به π^* می رسد و طی $|A|$ مرحله Policy iteration حتماً به π^* می رسیم که از هر سیاست دیگر بزرگتر است.

[مرجع: towards data science.com]

ب) طبق همان مدل Policy iteration که در بخش قبل معرفی کردیم، همان داریم برای a، argmax می گیریم، پس در هر مرحله از این iteration یک سیاست deterministic داریم که $a = \pi_i(s)$ و در مرحله آخر که به π^* می رسیم، $a = \pi^*(s)$ که از همه سیاست ها بهتر است و deterministic نیز هست.

PAPCO