به نام خدا

یادگیری تقویتی

نیم‌سال بهار ۴۰۲-۱۴۰۱

اساتید: دکتر رهبان، آقای حسنی

دانشکده‌ی مهندسی کامپیوتر

کوییز اول | ۲۵ بهمن | زمان تحویل: ۲۰ دقیقه

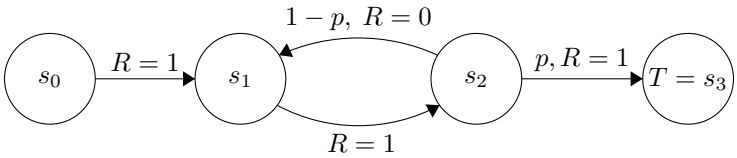
سوال ۱: (۲۰ نمره)

مشخص کنید کدام یک از گزاره های زیر برقرار است:

(آ) اضافه کردن مقدار ثابتی به تابع پاداش، policy بهینه را تغییر نمی‌دهد.

(ب) در هر MDP متناهی، حداقل یک policy بهینه deterministic وجود دارد.

(ج) زمانی که Value Iteration همگرا می‌شود، همواره به policy بهینه میل می‌کند.

سوال ۲: (۴۰ نمره)

در یک MDP به‌ازای یک policy مشخص مانند π، گذار زیر بین حالت‌ها و پاداش آن‌ها نمایش داده شده است. حالت s_3 حالت پایانی است، به این معنا که با وارد شدن به آن، اپیزود خاتمه میابد. فرض کنید $\gamma = 1$ است. معادلات bellman را برای value حالت‌ها نوشته، با حل آن‌ها $v_\pi(s_i)$ را به ازای $i = 0, 1, 2$ بدست آوردید. (توجه: $v_\pi(s_3) = 0$ فرض شود.)