به نام خدا

یادگیری تقویتی

نیم‌سال بهار ۴۰۲-۱۴۰۱

اساتید: دکتر رهبان، آقای حسنی

دانشکده‌ی مهندسی کامپیوتر

مدت: ۲۰ دقیقه | کوییز ۲ | تاریخ: ۱۴ اسفند

نام و نام خانوادگی : شماره دانشجویی:

سوال ۱: سوال توضیحی (۴۰ نمره)

(آ) یکی از روش‌های کاهش واریانس در روش‌های بهینه‌سازی سیاست، کم کردن یک baseline از پاداش تجمیعی است. توضیح دهید چگونه با این تفریق تغییری در متوسط گرادیان تابع هزینه نسبت به پارامترهای تابع سیاست رخ نمی‌دهد. (با نوشتن رابطه اثبات شود.)

(ب) چالش الگوریتم tabular Q-learning در محیط‌های با فضای حالت پیوسته چیست؟ توضیح دهید شبکه Deep Q-learning (DQN) چگونه قادر است بر این مشکل غلبه کند.

سوال ۲: الگوریتم REINFORCE (۶۰ نمره)

تصور کنید در یک محیط با فضای حالت پیوسته عامل تنها دو حرکت راست و چپ را دارد. همچنین در نظر بگیرید که تابع سیاست توسط شبکه عصبی‌ای با پارامتر θ محاسبه می‌شود: (تصور شود که مقدار $\phi(s,a)$ برای هر جفت حالت_کنش به صورت یک عدد به ما داده شده است.)

$$\pi_\theta(a|s) = \frac{\exp(\theta^\top \phi(s,a))}{\sum_{a'} \exp(\theta^\top \phi(s,a'))}$$

تابع پاداش نیز به صورت زیر بدست می‌آید:

$$R(s,a) = \begin{cases} -s^2 & ,a = Left \\ s^2 & ,a = Right \end{cases}$$

(آ) رابطه بروزرسانی تابع سیاست با الگوریتم REINFORCE را بنویسید.

(ب) تصور کنید مسیر زیر به طول T=3 توسط عامل طی شده و حال می‌خواهد تابع سیاست خود را بروز کند:

$$s_1 = 0.5, a_1 = Left$$

$$s_2 = -0.8, a_2 = Right$$

$$s_3 = 0.2, a_3 = Right$$

گرادیان تابع سیاست را با استفاده از مسیر طی‌شده و فرمول نوشته شده بخش (آ) محاسبه کنید. ضریب تخفیف (γ) را 0.9 در نظر بگیرید.