دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

**هوش مصنوعی**
**پاییز ۱۴۰۰**
استاد: محمدحسین رهبان
گردآورندگان: محمد محدی، حمیدرضا کامکاری
بررسی و بازبینی: محمدرضا یزدانی‌فر

پاسخ تمرین هفتم سری دوم — Reinforcement Learning — مهلت ارسال: ۱۶ دی

## سوالات (۱۰۰ نمره)

۱. (۱۰۰ نمره)

(آ) در این موارد بهتر است که از Q-value استفاده کنیم. زیرا اگر از V-value استفاده کنیم، برای استخراج سیاست از این مقادیر، باید به شکل زیر عمل کنیم:

$$V(s) = max_a R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s')$$

که باید روی تمام حالات $a$، بررسی کنیم کدام است که $V$ را بیشینه می‌کند. اما برای این کار، نیازمند دانستن $R$ و $P$ هستیم که در محیط ناشناخته، به آن‌ها دسترسی نداریم. اما اگر از Q-value استفاده کنیم، سیاست را می‌توانیم به صورت زیر استخراج کنیم:

$$\pi(s) = \text{argmax}_a Q(s,a)$$

(ب) در زمانی که از الگوریتم‌های online استفاده می‌کنیم، گاهی ممکن است که برخی از state ها و برخی action ها به اندازه‌ی کافی آزمایش نشده‌اند یا به کلی آزمایش نشده‌اند. بنابراین با وجود بدست آمدن یک سیاست بهینه، با احتمال کمی یکی از action ها را اجرا می‌کنیم تا محیط را به طور کامل explore کنیم. اگر هم این کار را انجام ندهیم، ممکن است که در یک ماکسیمم محلی گیر کنیم و به درستی تمام محیط را explore نکنیم.

(ج)

$$\begin{aligned}
V^{\pi}(s_1) &= \mathbb{E}_{a_1 \sim \pi}[Q^{\pi}(s_1,a_1)] \\
&\leq \mathbb{E}_{a_1 \sim \pi'}[Q^{\pi}(s_1,a_1)] \\
&= \mathbb{E}_{a_1 \sim \pi', a_2 \sim \pi}[R(s_1,a_1) + \gamma Q^{\pi}(s_2,a_2)] \\
&\leq \mathbb{E}_{a_1,a_2 \sim \pi'}[R(s_1,a_1) + \gamma Q^{\pi}(s_2,a_2)] \\
&= \mathbb{E}_{a_1,a_2 \sim \pi', a_3 \sim \pi}[R(s_1,a_1) + \gamma R(s_2,a_2) + \gamma^2 Q^{\pi}(s_3,a_3)] \\
&\leq \mathbb{E}_{a_1,a_2,\ldots \sim \pi'}[R(s_1,a_1) + \gamma R(s_2,a_2) + \gamma^2 R(s_3,a_3) + \ldots] = V^{\pi'}(s_1)
\end{aligned}$$

در نتیجه:

$$V^{\pi'}(s) \geq V^{\pi}(s)$$