تمرین سری اول

امیرحسین انتظاری

۲۲ آبان ۱۴۰۲

۱.۰ سوالات و پاسخ ها

۱. الف)

پاسخ:

هنگامی که یک تنظیم(adjustemnt) پاداش ثابت c به همه transition ها در MDP معرفی می‌شود، تابع مقدار بهینه جدید $V_۲$ می‌تواند به صورت زیر بیان شود:

$$V_۲(s) = V_۱(s) + \frac{c}{۱ - \gamma}$$

که در آن $V_۱(s)$ تابع مقدار بهینه اصلی و γ ضریب تخفیف $(discount factor)$ است.

استدلال پشت این عبارت این است که معرفی یک پاداش ثابت c برای همه انتقال‌ها به دلیل فرمول سری هندسی ارزش هر حالت را به میزان $\frac{c}{۱-\gamma}$ افزایش می‌دهد. ضریب تخفیف γ تأثیر این تنظیم را محدود می کند.

اکنون، با توجه به استراتژی بهینه ($\pi_۲$) در این سناریوی تغییر یافته، نکته مهم این است که توجه داشته باشیم که استراتژی$(policy)$ بهینه از تابع مقدار بهینه مشتق شده است. استراتژی $\pi_۲$ همچنان با انتخاب اعمالی که تابع مقدار بهینه جدید $V_۲$ را به حداکثر می‌رسانند تعیین می‌شود.

به طور کلی، اگر استراتژی بهینه $\pi_۱$ برای MDP اصلی شامل اقداماتی$(actions)$ نباشد که در مرز تصمیم‌گیری هستند (یعنی اقداماتی که به دلیل ساختار پاداش بی‌تفاوت بودند)، معرفی پاداش ثابت تنظیم c ممکن است استراتژی بهینه را تغییر ندهد. عامل همچنان اعمال یکسان را در هر حالت ترجیح می دهد زیرا تفاوت نسبی ارزش بین اقدامات در هر حالت یکسان باقی می ماند.

با این حال، اگر استراتژی اصلی شامل اعمالی در مرز تصمیم‌گیری باشد (جایی که عامل بین چندین اعمال بی‌تفاوت بود)، تعدیل پاداش ثابت c به طور بالقوه می‌تواند ترجیحات را تغییر دهد و منجر به تغییر در استراتژی بهینه شود. عامل کلیدی در اینجا این است که آیا تعدیل پاداش برای تغییر مقادیر نسبی اعمال در حالت‌های خاص کافی است و بر تصمیم‌گیری عامل تأثیر می‌گذارد. به عبارت دیگر استراتژی بهینه ممکن است بسته به ویژگی های MDP اصلی، تنظیم پاداش c و ویژگی های استراتژی بهینه اصلی $\pi_۱$ تغییر کند یا تغییر نکند.

۲. ب)

پاسخ:

اینکه آیا استراتژی بهینه پس از مقیاس بندی پاداش تغییر می کند یا ثابت می ماند، به ویژگی های خاص MDP اصلی و فاکتور مقیاس c بستگی دارد. انتخاب c که برای آن همه استراتژی ها بهینه هستند، مستلزم حفظ نظم نسبی مقادیر عمل $(action)$در حالت ها است. وقتی همه پاداش‌ها در MDP با یک ثابت $c \in R$ مقیاس‌بندی می‌شوند، تابع مقدار بهینه جدید $V_۲$ می‌تواند به صورت زیر بیان شود:

$$V_۲(s) = c \cdot V_۱(s).$$

استدلال پشت این عبارت این است که مقیاس دادن پاداش ها، تابع ارزش را به طور متناسب مقیاس می دهد.

حال در مورد استراتژی بهینه ($\pi_۲$) در این سناریو:

۱. استراتژی بهینه باقی بماند $\pi_۱$: - اگر استراتژی بهینه اولیه $\pi_۱$ به گونه‌ای باشد که تصمیم‌گیری عامل بر اساس مقادیر نسبی اقدامات باشد نه بزرگی مطلق آن‌ها، ممکن است استراتژی بهینه بدون تغییر باقی بماند. این به این دلیل است که مقیاس کردن همه پاداش ها توسط یک ثابت c ترتیب نسبی مقادیر عمل را در هر حالت تغییر نمی دهد.

- اگر $\pi_۱$ در MDP اصلی بهینه بود، اگر عامل به انتخاب اعمال در هر حالت بر اساس ترتیب نسبی مقادیر مقیاس شده آنها ادامه دهد، در MDP مقیاس شده بهینه می ماند.

- تابع مقدار حاصل در این مورد $V_۲(s) = c \cdot V_۱(s)$ است.

۲. استراتژی بهینه تغییر یابد: - اگر استراتژی بهینه اولیه $\pi_۱$ شامل اعمال در مرز تصمیم‌گیری می‌شد که در آن عامل بین چندین اعمال به دلیل مقادیر پاداش خاص بی‌تفاوت بود، مقیاس‌بندی پاداش‌ها ممکن است استراتژی بهینه را تغییر دهد.

- استراتژی بهینه ممکن است تغییر کند اگر مقیاس بندی ترتیب نسبی مقادیر عمل را در ایالت ها تغییر دهد و بر تصمیم گیری عامل تأثیر بگذارد.

- در این حالت، تابع مقدار حاصل ممکن است یک رابطه خطی ساده با تابع مقدار اصلی $V_۱$ نداشته باشد.

۳. انتخاب c برای بهینه بودن همه استراتژی ها: - انتخابی از c وجود دارد به طوری که همه استراتژی ها بهینه هستند اگر و فقط اگر MDP اصلی یک استراتژی بهینه منحصر به فرد $\pi_۱$ داشته باشد و فاکتور مقیاس c به گونه ای انتخاب شود که چنین باشد. ترتیب نسبی مقادیر عمل را در هیچ حالتی تغییر ندهید.

- اگر c به گونه ای انتخاب شود که $c > ۰$ و ترتیب نسبی مقادیر عملکرد در حالت ها پس از مقیاس بندی یکسان باقی بماند، همه سیاست ها، از جمله استراتژی بهینه اصلی $\pi_۱$ در MDP مقیاس شده بهینه باقی بماند.

۳. ج)
پاسخ:

وجود حالت های پایانی در MDP در واقع می تواند بر پاسخ به بخش (a) تأثیر بگذارد. در بخش (الف)، ما در مورد معرفی یک تنظیم پاداش ثابت c برای همه انتقال‌ها در یک MDP بدون حالت پایانی بحث کردیم. هنگامی که حالت‌های پایانی معرفی می‌شوند، پویایی MDP تغییر می‌کند و تأثیر تعدیل پاداش ممکن است متفاوت باشد.

بیایید مثالی از MDP را در نظر بگیریم که در آن پاسخ‌ها به بخش (الف) و قسمت (ج) متفاوت است:

** MDP مثال:** - حالات: $S = \{s_۱ s_۲ s_۳ T\}$، که در آن T یک حالت پایانی است. - عمل ها(actions): $A = \{a_۱, a_۲\}$. - انتقال و پاداش: - $s_۱ \xrightarrow{a_۱} s_۲$ با پاداش ۱ = $R(s_۱ a_۱)$. - $s_۲ \xrightarrow{a_۱} s_۳$ با پاداش ۲ = $R(s_۲ a_۱)$. - $s_۳ \xrightarrow{a_۱} T$ - با پاداش ۰ = $R(s_۳ a_۱)$. - $s_۱ \xrightarrow{a_۲} T$ - با پاداش ۱۰ = $R(s_۱, a_۲)$. - $s_۲ \xrightarrow{a_۲} T$ - با پاداش ۵ = $R(s_۲ a_۲)$. - $s_۳ \xrightarrow{a_۲} T$ - با پاداش ۱ = $R(s_۳ a_۲)$.

در این MDP، سیاست بهینه $\pi_۱$ بدون حالت های پایانه ممکن است شامل انجام عمل $a_۱$ در $s_۱$ و $s_۲$ و انجام عمل $a_۲$ در $s_۳$. تابع مقدار بهینه اصلی $V_۱$ بر اساس این سیاست تعیین می شود.

حال، اگر یک تنظیم پاداش ثابت c را برای همه انتقال‌ها در MDP، از جمله مواردی که به حالت‌های پایانی منتهی می‌شوند، معرفی کنیم، تابع مقدار بهینه جدید $V_۲$ تحت تأثیر قرار می‌گیرد. تأثیر روی استراتژی بهینه به ویژگی های تعدیل پاداش و عامل تخفیف بستگی دارد.

در مقابل، اگر حضور حالت‌های پایانی را در قسمت (c) در نظر بگیریم، مقیاس کردن همه پاداش‌ها با یک ثابت c همچنان بر تابع مقدار تأثیر می‌گذارد، اما استراتژی بهینه ممکن است تا زمانی که ترتیب نسبی عمل بدون تغییر باقی بماند. ارزش های درون ایالت ها حفظ می شود. با این حال، معرفی حالت‌های پایانه به MDP پیچیدگی می‌افزاید و تأثیر آن بر استراتژی بهینه باید در چارچوب مقیاس پاداش و $discount factor$ خاص تحلیل شود.