دانشگاه صنعتی امیرکبیر

دانشکده مهندسی کامپیوتر

«مبانی و کاربردهای هوش مصنوعی ترم پائیز ۱۴۰۴»

تمرین چهارم

در انجام تمرین‌ها به نکات زیر توجه فرمائید:

۱- مطابق قوانین دانشگاه هر نوع کپی‌برداری و اشتراک کار دانشجویان غیرمجاز است و پاسخ به تمرین‌ها باید به صورت انفرادی و بدون استفاده از ابزارهای هوش مصنوعی انجام شود، در صورت مشاهده چنین مواردی با طرفین شدیدا برخورد خواهد شد.

۲- پاسخ خود را در قالب یک فایل PDF به صورت تایپ‌شده و یا دستنویس (مرتب و خوانا) در سامانه کورسز آپلود نمائید.

۳- در صورت هر گونه سوال یا ابهام می‌توانید از طریق راه‌های ارتباطی گفته‌شده با تدریس‌یارهای طراح این تمرین در ارتباط باشید.

۴- توجه نمائید پاسخ تمرین‌ها تنها در صورت آپلود در سامانه کورسز پذیرفته خواهد شد و ارسال پاسخ از طریق ایمیل یا تلگرام بررسی نخواهد شد.

۵- فایل تمرین را با فرمت StudentID_AI_HW04.pdf تا ساعت ۲۳:۵۹ روز ۱۴۰۴/۰۹/۱۷ فقط در بخش مربوطه در سایت درس آپلود نمائید.

۶- دقت کنید که تمرین‌ها تاخیر مجاز ندارند و به ازای هر روز تاخیر، ۲۰٪ از نمره تمرین مربوطه کسر خواهد شد.

مسئله تصمیم‌گیری مارکوف و یادگیری تقویتی

سوال اول

تعداد n بندر در امتداد یک خط ساحلی وجود دارند که از 1 تا N شماره‌گذاری شده اند. تاجری در بندر یک زندگی می‌کند. این تاجر هرروز می‌تواند یکی از کار های زیر را انجام بدهد:

به شهر همسایه(شرق یا غرب) سفر کند،

یا در شهر فعلی بماند و تجارت کند.

اگر تاجر تصمیم بگیرد از بندر i سفر کند، با احتمال Pi با موفقیت به بندر بعدی می‌رسد، اما با احتمال 1- Pi به طوفان بر می‌خورد و سفر نمی‌کند.

اگر تصمیم بگیرد در شهر i بماند و تجارت کند، پاداشی بزرگتر از صفر برابر با ri دریافت می‌کند. در مقابل، روزی که برای سفر سفر صرف می‌شود، پاداشی ندارد. چه به بندر بعدی برسد و چه نرسد.

الف) اگر برای همه بندر ها ri = 1، Pi = 1 و ضریب تخفیف برابر با 0.5 باشد، مقدار $V_{stay}(1)$ یعنی مقدار بودن در بندر 1 تحت سیاستی که همیشه ماندن را انتخاب کند، چقدر است ؟

ب) اگر برای همه i ها، ri = 1 و Pi = 1 و ضریب تخفیف برابر با 0.5 باشد، مقدار بهینه $V^*(1)$ چقدر است ؟

پ) اگر همه ri ها و Pi ها اعداد مثبتی باشند و ضریب تخفیفی نداشته باشیم، سیاست بهینه را توصیف کنید.

ت) فرض کنید الگوریتم value iteration را اجرا می‌کنید. می‌دانید $V_k(s)$ مقدار حالت s بعد از k بار اجرای الگوریتم مورد نظر است و مقدار اولیه همه حالت ها برابر صفر است. اگر مقدار بهینه در بندر اول مثبت باشد، بزرگترین k که در آن $V_k(1)$ همچنان صفر است چیست ؟

ث) اگر همه ri ها و Pi ها مثبت باشند، بزرگترین k که در آن $V_k(s)$ برای بعضی از حالات s همچنان صفر است چیست ؟ (فرض کنید مقدار اولیه حالات قبل از اجرای الگوریتم تکرار مقدار صفر است.)

ج) فرض کنید که حالات زیر را تجربه کردیم. تجربه ها به صورت یک چهار تایی مرتب هستند.

(s=1, a=stay, r=4, s'=1)

(s=1, a=east, r=2, s'=2)

(s=2, a=stay, r=6, s'=2)

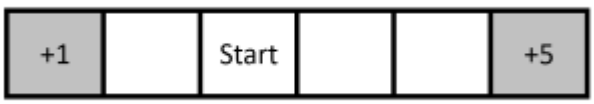
(s=2, a=west, r=0, s'=1)

(s=1, a=stay, r=4, s=1)

بعد از هر تجربه، مقادیر Q(1,Stay)، Q(1,east)، Q(2,west) و Q(2,stay) را حساب کنید. فرض کنید نرخ یادگیری 0.5 و ضریب تخفیف 1 است و مقدار اولیه مقادیر مورد نظر صفر است.

سوال دوم

محیط gridWord ساده‌ی یک بعدی زیر را در نظر بگیرید که در آن یک عامل از خانه Start شروع کرده و در هر کدام از خانه های سفید توانایی انتخاب کنش های چپ و راست را دارد. در خانه های خاکستری رنگ، تنها انتخاب ممکن خروج از بازی است و به هنگام خروج از هر کدام، پاداشی دریافت می‌شود که در خانه مربوطه نوشته شده است. به هنگام ترک خانه های سفید پاداشی دریافت نمی‌شود. فرض کنید ضریب تخفیف برابر یک است. به سوالات زیر پاسخ دهید.



الف) میزان ارزش بهینه حالت Start چقدر است؟

ب) با فرض اجرای الگوریتم value iteration بعد از چند گام k، خواهیم داشت $V^*(Start) = V_k(Start)$ ؟ آیا ممکن است که این دو مقدار مساوی نشوند ؟

ج) فرض کنید ضریب تخفیف 0.8 باشد، مقدار بهینه حالت Start چقدر است ؟

د) ضریب تخفیف چه مقادیری می‌تواند داشته باشد تا سیاست بهینه در حالت Start حرکت به سمت چپ باشد ؟

ه) دوباره فرض کنید ضریب تخفیف 1 است اما اعمال ممکن در خانه های سفید با احتمال 0.8 منجر به گذار به خانه های مجاور می‌شوند. در غیر این صورت به سمت خانه مخالف حرکت می‌کند. حال مقدار بهینه حالت شروع چیست ؟

و) در این شرایط اولین گام k در الگوریتم value iteration که در آن $V_k(Start)$ غیر صفر می‌شود، چند است ؟

ز) بعد از چند گام k، خواهیم داشت $V^*(Start) = V_k(Start)$ ؟ آیا ممکن است که این دو مقدار مساوی نشوند ؟

سوال سوم

صحیح یا غلط بودن سوالات زیر را بررسی کنید. برای هر مورد یک توضیح کوتاه نیز بنویسید.

الف) روش یادگیری تقویتی مبتنی بر مدل، علاوه بر value یا تابع policy، مدلی از محیط، از جمله احتمالات و پاداش های انتقال را یاد می‌گیرد.

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

ج) زمانی که value iteration همگرا می‌شود، همواره به policy بهینه میل می‌کند.

د) وقتی ضریب نخفیف نزدیک به صفر باشد، عامل تمایل بیشتری به گرفتن پاداش های بلند مدت خواهد داشت.

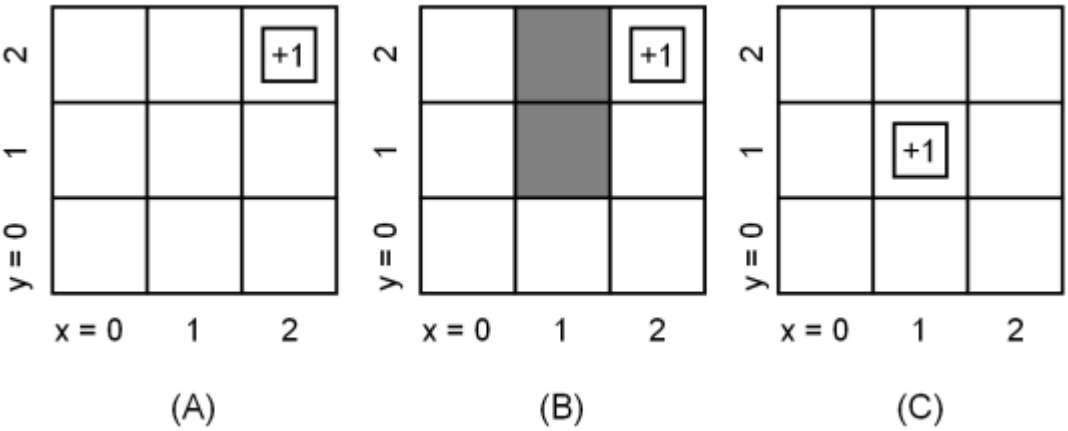
ه) Q-Learning یک الگوریتم model-free است.

و) در ε-Greedy اگر ε ثابت بماند، عامل نمی‌تواند سیاست بهینه را یافت کند.

ز) همگرایی یادگیری Q کاملا به سیاستی که دنبال می‌کند وابسته می‌باشد.

سوال چهارم

در gridWorld های زیر، یک عامل وجود دارد و این عامل می‌تواند چهار عمل east، north، west و south را انتخاب کند. در اینجا نویزی وجود ندارد مگر اینکه جهت حرکت عامل به سمت خارج جدول یا خانه های خاکستری باشد که در این صورت عامل در جای خود می ماند. در هر قسمت وقتی عامل وارد خانه جعبه‌ای شکل می‌شود، مجاز است تا با انجام عمل exit از بازی خارج بشود و پاداش مشخص شده در خانه را دریافت کند. به غیر از این بقیه پاداش ها صفر هستند و ضریب تخفیف برابر 0.5 است.



الف) مقادیر بهینه برای جدول A را بیابید.

ب) سیاست بهینه در جدول B را مشخص کنید.

پ) فرض کنید برای هر حالت غیر پایانی s مجموعه ای از ویژگی هایی با مقدار حقیقی $f_i(s)$ داریم و می‌خواهیم مقادیر بهینه هر حالت را با تقریب خطی زیر تخمین بزنیم:

$$\sum_i w_i f_i(s) = V(s)$$

اگر ویژگی های ما $f_1(x,y) = x$ و $f_2(x,y) = y$ باشند، وزن ها را طوری تعیین کنید که در جدول A سیاست استخراج شده بهینه باشد.

ت) آیا مقادیر بهینه واقعی را می توان با استفاده از ویژگی های قسمت قبل بدست آورد؟ دلیل بیاورید.

ث) برای هر کدام از مجموعه ویژگی های زیر، در کدام جدول‌ها سیاست بهینه بدست می‌آید؟ دلیل بیاورید.

۱) $f_1(x,y) = x$ و $f_2(x,y) = y$

۲) $f_1(x,y) = (x-1)^2$، $f_2(x,y) = (y-1)^2$ و $f_3(x,y) = 1$

۳) برای هر (i,j) که مختصات یک حالت هستند، $f_{i,j}(x,y) = 1$ اگر (x,y) = (i,j) باشد در غیر این صورت صفر است.

سوال پنجم

در این سؤال، قصد داریم یک مدل تصمیم‌گیری مارکوف (MDP) را برای انتخاب وسیله‌ی بازی در یک شهربازی طراحی و تحلیل کنیم. فرض کنید شما به یک شهربازی رفته‌اید که در آن وسایل متنوعی وجود دارد. وظیفه‌ی شما انتخاب هوشمندانه‌ی وسایل با هدف بیشینه‌سازی پاداش کلی است.

در ابتدا حال شما خوب است و از وسایل مختلف پاداش‌هایی دریافت می‌کنید که بسته به نوع وسیله می‌توانند متفاوت باشند. با این حال، احتمال دارد برخی از وسایل باعث بیمار شدن شما شوند. اگر در حالتی که بیمار هستید، به استفاده از وسایل ادامه دهید، ممکن است با گذشت زمان حالتان دوباره خوب شود، ولی در این وضعیت پاداش‌هایی که دریافت می‌کنید کاهش می‌یابد و حتی ممکن است منفی شود.

شما قبلاً این وسایل را امتحان نکرده‌اید، بنابراین نمی‌دانید هر وسیله چه پاداشی در حالت سالم یا بیمار به شما می‌دهد. همچنین اطلاعی از احتمال بیمار شدن یا بهبودی پس از استفاده از وسایل مختلف ندارید. با این وجود، اطلاعاتی درباره‌ی ویژگی‌های وسایل در اختیار دارید:

اقدام/وسیله	نوع	مدت انتظار	سرعت
گردباد رنگین‌کمان	ترن هوایی	طولانی	سریع
شهاب‌سنگ آبی	ترن هوایی	کوتاه	کند
سقوط آتشفشانی	برج سقوط	کوتاه	سریع
پرواز مه‌آلود	تاب آونگی	کوتاه	کند
تَرک پارک	خروج	کوتاه	کند

ما این مسئله را به‌صورت یک MDP با دو وضعیت «سالم» و «بیمار» مدل می‌کنیم. هر وسیله‌ی بازی یک اقدام (action) محسوب می‌شود. اقدام «ترک پارک» نیز فرآیند تصمیم‌گیری را متوقف می‌کند. با انتخاب هر وسیله، ممکن است وضعیت جسمانی شما ثابت بماند یا تغییر کند.

برای محاسبه‌ی مقدار Q، از یک تخمین تقریبی مبتنی بر ویژگی‌ها (feature-based approximation) استفاده می‌کنیم. ویژگی‌ها و وزن‌های اولیه‌ی متناظرشان به صورت زیر هستند:

وزن اولیه	ویژگی‌ها
$w_0 = 1$	$f_0(state, action) = 1 \rightarrow$ ویژگی بایاس (همیشه برابر ۱)
$w_1 = 2$	$f_1(state, action) = 1$ اگر نوع وسیله «ترن هوایی» باشد، وگرنه صفر
$w_2 = 1$	$f_2(state, action) = 1$ اگر مدت انتظار «کوتاه» باشد، وگرنه صفر
$w_3 = 0.5$	$f_3(state, action) = 1$ اگر سرعت «سریع» باشد، وگرنه صفر

(الف) مقدار تقریبی Q را برای وضعیت «سالم» و اقدام «گردباد رنگین‌کمان» محاسبه کنید:

یعنی:

$$Q\left(\text{سالم،گردباد رنگین‌کمان}\right) = \text{؟}$$

(ب) با استفاده از یک نمونه تجربه به شکل زیر:

$$\left(\text{سالم ، گردباد رنگین کمان ، بیمار} , -10.5\right)$$

و با استفاده از نرخ یادگیری $\alpha = 0.5$ و ضریب تخفیف $\gamma = 0.5$، مقدار وزن‌های جدید را پس از انجام یک مرحله به‌روزرسانی Q-learning محاسبه کنید.

(پ) آیا مقادیر Q مربوط به وضعیت «سالم» با مقادیر Q در وضعیت «بیمار» برای هر وسیله یکسان هستند؟

به عبارت دیگر، آیا برای هر اقدام داریم:

$$Q\left(\text{سالم،وسیله}\right) = Q\left(\text{وسیله بیمار،همان}\right)$$

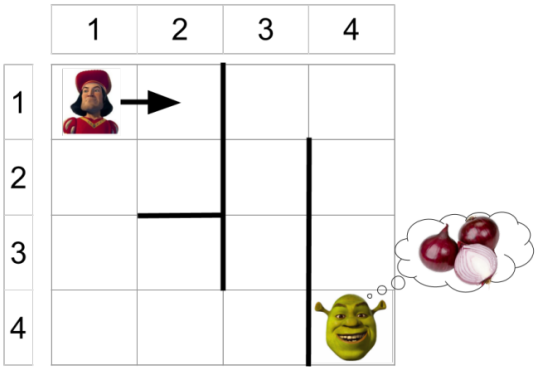
در صورت مثبت یا منفی بودن پاسخ، دلیل آن را به‌روشنی توضیح دهید. تمرکز شما باید بر نقش ویژگی‌های تعریف‌شده در تعیین تفاوت یا شباهت این مقادیر باشد.

(ت) فرض کنید هنوز از وزن‌های اولیه استفاده می‌شود. اگر در وضعیت «سالم» از یک سیاست ϵ-greedy استفاده شود، کدام وسیله با چه احتمالی انتخاب خواهد شد؟

اگر چندین وسیله قابل انتخاب باشند، تمام آن‌ها و احتمال انتخابشان را ذکر کنید.

سوال ششم

در این سؤال، قصد داریم با استفاده از مفهوم MDP (فرایند تصمیم‌گیری مارکوف)، یک محیط شبیه‌سازی‌شده را مدل کنیم که در آن شخصیت اصلی (Lord Farquaad) می‌خواهد به هدف مشخصی برسد.



تعریف محیط MDP:

شخصیت Farquaad در یک شبکه مربعی ۴×۴ حرکت می‌کند. در هر لحظه، او نه‌تنها در یک موقعیت مکانی قرار دارد، بلکه یک جهت‌گیری (جهت نگاه) نیز دارد. به همین دلیل، فضای حالت‌ها به‌صورت سه‌تایی تعریف می‌شود:

مختصات سطری (row)

مختصات ستونی (column)

جهت‌گیری که یکی از چهار گزینه {شمال، شرق، جنوب، غرب} است.

Farquaad در آغاز بازی در وضعیت (ردیف ۱، ستون ۱، جهت شرق) قرار دارد.

اقدامات (Actions):

سه اقدام مجاز برای Farquaad تعریف شده است:

R: چرخش به راست (بدون تغییر موقعیت)

L: چرخش به چپ (بدون تغییر موقعیت)

M: حرکت رو به جلو در راستای جهت فعلی (اگر دیوار مانع نباشد)

حرکت به‌سمت دیوار منجر به باقی ماندن در همان وضعیت می‌شود، اما همچنان به‌عنوان اقدام ثبت می‌شود.

تابع پاداش (Reward Function):

Farquaad تنها زمانی پاداش دریافت می‌کند که وارد خانه‌ی خاصی شود که مکان Shrek (باتلاق) است. این پاداش برابر با ۵ است. در سایر خانه‌ها پاداش صفر است.

تابع انتقال (Transition Function):

محیط کاملاً قطعی (Deterministic) است. یعنی با اجرای یک اقدام خاص از یک حالت مشخص، همیشه به یک حالت مشخص می‌رویم (اگر قابل دسترس باشد). در غیر این صورت، وضعیت تغییری نمی‌کند.

---

۱. اندازه‌ی فضای حالت‌ها $|S|$ و فضای اقدامات $|A|$ را به‌دست آورید.
(در اینجا منظور از $|S|$ تعداد کل وضعیت‌های ممکن و از $|A|$ تعداد کل اقدامات مجاز است.)

۲. چرا به این مدل یک «فرایند مارکوف» گفته می‌شود؟
راهنما: به وابستگی احتمال انتقال بین وضعیت‌ها توجه کنید.

۳. مقادیر احتمالات انتقال زیر را مشخص کنید. منظور از علامت | در هر عبارت این است که «با داشتن شرط سمت راست، احتمال رسیدن به وضعیت سمت چپ چقدر است». به‌عبارت دیگر:

$$p(s' \mid s, a)$$

به معنی احتمال رفتن از حالت s به حالت s' با اقدام a است.

عبارات زیر را مقداردهی کنید:

$$p\left(\left(1,1, \text{شمال}\right) \mid \left(1,1, \text{شمال}\right), M\right) = \text{؟}$$

$$p\left(\left(1,1, \text{شمال}\right) \mid \left(1,1, \text{شرق}\right), L\right) = \text{؟}$$

$$p\left(\left(2,1, \text{جنوب}\right) \mid \left(1,1, \text{جنوب}\right), M\right) = \text{؟}$$

$$p\left(\left(2,1, \text{شرق}\right) \mid \left(1,1, \text{جنوب}\right), M\right) = \text{؟}$$

۴. فرض کنید Farquaad در وضعیت شروع (۱, ۱, شرق) قرار دارد و ضریب تخفیف $\gamma = 0.5$ است. مقدار پاداش تخفیف‌یافته‌ی مورد انتظار را در دو حالت زیر محاسبه کنید:

اگر اقدام اولیه R (چرخش به راست) باشد.

اگر اقدام اولیه L (چرخش به چپ) باشد.

(توجه: فرض کنید سیاست بهینه پس از اقدام اولیه دنبال می‌شود.)

(توجه :در تمام سؤالات بعدی، مقدار γ (ضریب تخفیف) برابر با ۰.۵ در نظر گرفته شود مگر اینکه مقدار تخفیف به صورت ضمنی در صورت سوال تغییر کرده باشد.)

۵. اگر جهت‌گیری در تمام حالات برابر با شرق باشد، در هر موقعیت (ردیف و ستون)، اقدام بهینه چیست؟ (اگر چند گزینه وجود دارد، یکی را انتخاب کنید.)

۶. فرض کنید شخصیتی به نام Vector پیشنهاد می‌دهد مقدار γ را از ۰.۵ به ۰.۹ افزایش دهیم. آیا این تغییر باعث تغییر در سیاست بهینه می‌شود؟ چرا؟

۷. حال فرض کنید تابع پاداش به شکل زیر تغییر کند:
در صورتی که Farquaad وارد باتلاق شود: $R(s,a) = 0$
در غیر این صورت: $R(s,a) = -1$
آیا این تغییر منجر به تغییر سیاست بهینه خواهد شد؟ دلیل بیاورید.

۸. شخصیت Elsa وارد شده و زمین را یخی کرده است. در این حالت، محیط تصادفی شده است. به این معنا که در اجرای اقدام M، با احتمال ۰.۲ Farquaad دو خانه به جلو سر می‌خورد. اگر وضعیت اولیه Farquaad برابر با $s = (2,4, \text{جنوب})$ باشد، مقدار پاداش تخفیف‌یافته‌ی مورد انتظار را از این حالت محاسبه کنید.

سوال هفتم (امتیازی)

در بیشتر الگوریتم‌های مرتبط با MDP، هدف یافتن $V^*(s)$ است، یعنی بیشینه پاداش انتظاری که یک عامل می‌تواند از حالت s با پیروی از سیاستی بهینه کسب کند. این مقدار، به‌صورت مجموع تخفیف‌یافته‌ای از پاداش‌ها تعریف می‌شود و توسط رابطه‌ی مشهور بلمن مشخص می‌گردد:

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

اما در این سؤال، کیفیت یک سیاست را نه با امید ریاضی پاداش، بلکه با بدترین پاداش ممکن که می‌تواند اتفاق بیفتد، می‌سنجیم.

دقیق‌تر، اگر از حالت s شروع کنیم و از یک سیاست خاص π پیروی کنیم، $L^\pi(s)$ برابر است با حداقل پاداش تخفیف‌یافته‌ای که ممکن است در یک مسیر ممکن اتفاق بیفتد.

در ادامه، تعریف می‌کنیم:

$$L^*(s) = \max_\pi L^\pi(s)$$

یعنی بزرگ‌ترین مقدار تضمینی از کمترین پاداشی که می‌توان با یک سیاست مناسب از حالت s به‌دست آورد.

برای ساده‌سازی، فرض می‌کنیم:

$$C(s,a) = \{s' \mid T(s,a,s') > 0\}$$

یعنی مجموعه‌ای از حالت‌هایی که ممکن است عامل پس از اجرای اقدام a از حالت s به آن منتقل شود (با احتمال غیر صفر).

(الف) $L^*(s)$ را به‌صورت بازگشتی مشابه معادله بهینگی بلمن بنویسید.

(ب) الگوریتم تکرار مقدار (Iteration Value) برای محاسبه $V^*(s)$ معمولاً طبق رابطه زیر عمل می‌کند:

$$V_{i+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_i(s')]$$

اکنون، نسخه‌ای مشابه از این به‌روزرسانی را برای تخمین $L^*(s)$ تعریف کنید، به‌طوری که از $L_i(s)$ برای محاسبه $L_{i+1}(s)$ استفاده شود.

(پ) فرض کنید:

تابع پاداش فقط به حالت بستگی دارد، یعنی $R(s,a,s') = R(s)$

برای تمامی حالات $R(s) \geq 0$

در حالت عادی، تابع Q برای پاداش انتظاری به‌صورت زیر تعریف می‌شود:

$$Q^*(s,a) = R(s) + \gamma \sum_{s\prime} T(s,a,s') \max_{a\prime} Q^*(s',a')$$

اکنون فرض کنید $M(s,a)$ نشان‌دهنده کمترین پاداش تضمینی است که عامل می‌تواند پس از اجرای اقدام a در حالت s به‌دست آورد و سپس بهینه عمل کند. رابطه بازگشتی برای $M^*(s,a)$ را مشابه با معادله بالا بنویسید.

(ت) در یادگیری Q کلاسیک، مقدار $Q(s,a)$ با استفاده از رابطه زیر به‌روزرسانی می‌شود:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left[R(s) + \gamma \max_{a\prime} Q(s',a')\right]$$

اکنون، چهار گزینه‌ی پیشنهادی برای یادگیری $M(s,a)$ ارائه شده‌اند. فرض کنید که عامل همه زوج‌های (s,a) را بی‌نهایت بار تجربه می‌کند.

از بین موارد زیر، مشخص کنید کدام به‌روزرسانی تضمین می‌کند که $M(s,a)$ به مقدار بهینه $M^*(s,a)$ همگرا خواهد شد.

اگر بیش از یکی معتبر است، آن‌که سریع‌تر همگرا می‌شود را مشخص کنید.

گزینه‌ها:

(i)

$$M(s,a) \leftarrow (1-\alpha)M(s,a) + \alpha\left[R(s) + \gamma \sum_{s\prime \in C(s,a)} \max_{a\prime} M(s',a')\right]$$

(ii)

$$M(s,a) \leftarrow (1-\alpha)M(s,a) + \alpha\left[R(s) + \gamma \min_{s\prime \in C(s,a)} \max_{a\prime} M(s',a')\right]$$

(iii)

$$M(s,a) \leftarrow R(s) + \gamma \min_{s\prime \in C(s,a)} \max_{a\prime} M(s',a')$$

(iv)

$$M(s,a) \leftarrow (1-\alpha)M(s,a) + \alpha \cdot \min\{M(s,a), R(s) + \gamma \max_{a\prime} M(s',a')\}$$