تحلیل یادگیری تقویتی Q برای مسیریابی گروهی بلادرنگ در محیط پویا با اعتبارسنجی مونت‌کارلو و پاداش تطبیقی

تاریخ دریافت: ۱۴۰۴/۱۱/۱۱
تاریخ پذیرش: ۱۴۰۵/۰۵/۰۴

ایمان شفیعی‌نژاد*[۱]، فائزه شکری[۲]
۱ استادیار پژوهشگاه هوافضا، وزارت علوم، تحقیقات و فناوری، تهران، ایران، shafieenejad@ari.ac.ir
۲ دانشجوی دکتری، پژوهشگاه هوافضا، وزارت علوم، تحقیقات و فناوری، تهران، ایران

چکیده

هدایت بلادرنگ گروه‌های رباتیک در محیط‌های پویا دارای موانع ثابت و متحرک، یکی از چالش‌های پیچیده در زمینه برنامه ریزی رباتیک خودران است. این پژوهش یک چارچوب پیشرفته یادگیری تقویتی ارائه می‌دهد که عملکرد دو الگوریتم یادگیری Q و یادگیری Q دوگانه را در هدایت گروهی ربات های هوایی مقایسه و بهبود می دهد. مطالعه حاضر از یک شبیه سازی شامل ۱۰ ربات هوایی، ۱۲ مانع ثابت با پراکندگی هوشمند و ۲ مانع متحرک با الگوی حرکت تصادفی است. دو ساختار حرکتی سیستم ۴ حرکتی و سیستم ۶ حرکتی (حرکت‌های مورب) بهره برده است. ساز و کارهای پاداش نوآورانه شامل پاداش پیشرفت (بر پایه کاهش فاصله اقلیدسی از هدف) و پاداش موفقیت گروهی پیاده سازی شد. نتایج نشان داد یادگیری Q دوگانه با ساختار ۶ حرکتی بالاترین عملکرد را با دقت موقعیتی ۷۸,۲ درصد و میانگین پاداش ۱۵۶,۱۸ ارائه داد. چارچوب مبتنی بر روش مونت‌کارلو ارائه شده، روشی قابل‌اطمینان برای ارزیابی الگوریتم‌های یادگیری تقویتی در سناریوهای پیچیده فراهم می‌کند. این چارچوب می‌تواند در کاربردهایی مانند تحویل خودران و عملیات جستجو و نجات به کار رود.

واژه‌های کلیدی: *مسیریابی بلادرنگ، رباتیک گروهی، یادگیری تقویتی Q دوگانه، موانع متحرک، شبیه‌سازی مونت‌کارلو*

A comparative analysis of q-learning variants for real-time swarm robotic path planning in dynamic environments with static and mobile obstacles using monte carlo validation and an adaptive reward mechanism

Iman Shafieenejhad [1], Faezeh Shokri [2]

1- Assistant Professor, Aerospace Research Institute, Ministry of Science, Research and Technology, Tehran, Iran
2- Ph.D. Student, Aerospace Research Institute, Ministry of Science, Research and Technology, Tehran, Iran

Abstract

Real-time guidance of robotic swarms in dynamic environments with static and mobile obstacles is a complex challenge in autonomous robotic planning. This research presents an advanced reinforcement learning framework comparing Q-learning and Double Q-learning for cooperative guidance of aerial robot swarms, using a simulation with 10 aerial robots, 12 static obstacles, and 2 randomly moving mobile obstacles. Two motion structures were utilized: a 4-directional system (cardinal directions) and a 6-directional system (with diagonal movements). Innovative reward mechanisms, including progress reward (based on reduced Euclidean distance to target) and group success reward, were implemented. Results demonstrated that Double Q-learning with the 6-directional structure achieved the highest performance, with positional accuracy of 78.2% and average reward of 156.18. The Monte Carlo-based framework provides a reliable evaluation method for reinforcement learning algorithms in complex scenarios, applicable to autonomous delivery and search and rescue operations.

Keywords: *Real-Time Path Planning, Swarm Robotics, Double Q-Learning, Mobile Obstacles, Monte Carlo Simulation*

۱. مقدمه

هدایت بلادرنگ گروه‌های رباتیک در محیط‌های پویا و ناشناخته، به عنوان یکی از چالش‌های پیش‌رو در حوزه رباتیک خودران و سیستم‌های چند عامله مطرح است. توانایی ربات های هوایی برای حرکت هماهنگ در فضاهای پیچیده دارای موانع ثابت و متحرک، کاربردهای گسترده‌ای در عملیات‌های امداد و نجات، پایش محیطی، کشاورزی دقیق و لجستیک هوشمند دارد. با این حال، حضور موانع پویا و نیاز به تصمیم‌گیری بلادرنگ، این مسئله را از یک مسیریابی ساده به یک مسئله مسیریابی پویا و چند معیاره تبدیل کرده است[۱-۳]. مسئله اصلی در هدایت گروهی ربات‌ها، هماهنگی میان عوامل برای دستیابی به هدف مشترک، در عین پرهیز از برخورد با موانع و یکدیگر است. محیط‌های عملیاتی اغلب شامل موانع ثابت (مانند ساختمان‌ها، درختان) و موانع متحرک (مانند حیوانات، خودروها، یا حتی سایر ربات‌ها) هستند که پیش‌بینی موقعیت آن‌ها را دشوار می‌سازد. این پویایی، روش‌های برنامه‌ریزی مسیر کلاسیک را که بر پایه نقشه‌های ایستا طراحی شده‌اند، با محدودیت مواجه می‌کند و نیازمند رویکردهای هوشمند و سازگار شونده است[۴-۶].

در میان روش‌های نوین و هوشمند مسیریابی بلادرنگ، یادگیری تقویتی به عنوان یک روش قدرتمند برای یادگیری رفتار بهینه از طریق تعامل مستقیم با محیط، توجه بسیاری را به خود جلب کرده است. در این روش، عامل با دریافت پاداش یا جریمه از محیط، به تدریج سیاست بهینه حرکت را فرا می‌گیرد. این ویژگی، یادگیری تقویتی را به ویژه برای محیط‌های پویا و غیرقطعی که مدل دقیق آن‌ها در دسترس نیست، مناسب ساخته است. در این روش، عامل نیازی به دانش پیشین کامل از محیط ندارد و می‌تواند در حین عمل، دانش خود را به روز کند[۷-۹]. در میان روش‌های یادگیری تقویتی[۱]، الگوریتم‌های خانواده یادگیری Q [۲] به دلیل سادگی مفهومی و کارایی عملی، جایگاه ویژه‌ای دارند. با این حال یادگیری Q استاندارد ممکن است به دلیل پدیده برآورد بیش از حد مقادیر Q، منجر به یادگیری سیاست‌های ناکارآمد شود. برای غلبه بر این نقیصه، الگوریتم یادگیری Q دوگانه[۳] ارائه شده است که با استفاده از دو جدول Q مجزا، عمل به تخمین بهینه‌تر مقادیر می‌کند و پایداری یادگیری را افزایش می‌دهد. بررسی مقایسه‌ای این دو الگوریتم در محیط‌های پیچیده گروهی، موضوعی است که نیاز به بررسی بیشتر دارد که در پژوهش حاضر به آن پرداخته شده است[۱۰-۱۳].

در ادامه به مرور پیشینه پژوهش با رویکرد مسیریابی بلادرنگ هوشمند پرداخته می شود. واتکینز و دایان الگوریتم یادگیری Q را به عنوان یک روش ساده و افزایشی برای یادگیری بهینه تصمیم‌ها در محیط‌های مارکوفی کنترل شده معرفی کرده‌اند. این الگوریتم با بهبود پی درپی مقادیر Q ، کیفیت اعمال مختلف را ارزیابی می‌کند. همچنین به‌روزرسانی چند مقدار Q به‌طور همزمان ارائه شده است[۱۴]. هاسلت و همکاران به بررسی مشکل بیش برآورد مقادیر عمل[۴] در یادگیری Q می‌پردازد. پژوهشگران نشان می دهند که ایده یادگیری Q عمیق[۵]، که ابتدا در محیط های جدولی معرفی شده بود، می تواند برای تقریب توابع در مقیاس بزرگ تعمیم یابد. با ارائه یک سازگارسازی خاص برایDQN، الگوریتم حاصل نه تنها بیش برآورد را

کاهش می‌دهد بلکه باعث بهبود عملکرد می‌شود[۱۴]. نگوِین و همکاران الگوریتم‌های On-policyو Off-policy یادگیری Q را برای کنترل ربات‌های پاندول وارونه دوچرخ (TWIP) ارائه کردند. در این روش‌ها، داده‌ها بدون نیاز به مدل جمع‌آوری می‌شوند و سیستم به دو زیر سیستم تقسیم شده تا حجم داده‌ها کاهش یابد و سرعت محاسبات افزایش یابد. نتایج شبیه‌سازی نشان داد که تکنیک‌های Off-policy با کارایی داده‌ای بالاتر و دقت بیشتر، عملکرد بهتری در کنترل سیستم‌های TWIP دارند[۱۱]. گئو و همکاران الگوریتم Q-MA مبتنی بر یادگیری Q و الهام از رفتار اجتماعی حیوانات (ممیتیک[۶]) را برای حل مسئله مسیریابی مشترک وسایل نقلیه در برداشت و حمل و نقل کشاورزی هوشمند ارائه کردند. الگوریتم با استراتژی مقدار دهی اولیه ترکیبی، سه عملگر تقاطع تخصصی و جستجوی محلی تطبیقی هدایت شده توسط یادگیریQ، جمعیت متنوعی تولید کرده و راه حل‌های با کیفیت بالا ارائه می‌دهد. نتایج آزمایش‌ها نشان داد که Q-MA عملکرد بهتری نسبت به الگوریتم‌های مرجع دارد و مؤلفه‌های الگوریتم به بهبود همگرایی و کیفیت راه‌حل کمک می‌کنند[۱۵]. دوان و همکاران الگوریتم MOQLCPSO مبتنی بر یادگیری Q و PSO چند هدفه را برای مسیریابی ربات‌ها در زمین‌های ناهموار ارائه کردند. این الگوریتم با سازگاری پویا پارامترها و عملگر تقاطع استراتژیک مسیرهای بهینه با حداقل طول و ناهمواری زمین تولید می‌کند. نتایج نشان داد که MOQLCPSO نسبت به روش‌های پیشرفته، دقت، تنوع جمعیت و کارایی بهینه‌بالاتری دارد و برای کاربردهای چند هدفه مسیریابی و تخصیص منابع مناسب است[۱۶]. لی و همکاران الگوریتم تکرار حریصانه‌ای مبتنی بر یادگیری Q را برای زمان‌بندی گروهی چند سناریویی پیشنهاد کردند. در این روش، یادگیری Q برای بهبود انتخاب جستجوی محلی و افزایش تنوع راه‌حل‌ها به کار رفت. نتایج نشان داد الگوریتم پیشنهادی از نظر کارایی و پایداری از روش‌های موجود برتر است[۱۷]. جیوفرید و همکاران یک مطالعه تجربی بر روی الگوریتم‌های یادگیری چندعاملی برای مسیریابی هم‌زمان در محیط‌های انبار انجام دادند که مسیرهای بدون برخورد برای چندین عامل را محاسبه می‌کند. این پژوهش الگوریتم‌های مبتنی بر یادگیری را با روش‌های سنتی جست‌وجو و برنامه‌ریزی مقایسه کرده و کاربردپذیری آن‌ها در محیط‌های واقعی را ارزیابی می‌کند. نتایج نشان داد برخی الگوریتم‌های یادگیری چندعاملی عملکردی نزدیک به روش‌های سنتی با کیفیت مسیر مشابه و تلاش محاسباتی کمتر دارند، که نشان‌دهنده پتانسیل بالای آن‌ها برای کاربردهای عملی است[۱۸].

ژو و همکاران یک الگوریتم یادگیری Q بهینه‌شده برای برنامه‌ریزی مسیر محلی ربات‌های متحرک پیشنهاد کردند. آن‌ها با معرفی روش‌های جدید مقداردهی اولیه جدول Q، سیاست انتخاب عمل و تابع پاداش و نیز استفاده از روش RMSprop برای تنظیم پویای نرخ یادگیری، به بهبود همگرایی و جلوگیری از افتادن در بهینه‌های محلی پرداختند. نتایج شبیه‌سازی در محیط‌های مارپیچ با پیچیدگی‌های مختلف نشان داد که الگوریتم پیشنهادی در معیارهایی مانند تعداد گام‌ها، پاداش اکتشاف و زمان اجرا نسبت به روش‌های موجود برتری دارد. این پژوهش مسیر را برای توسعه الگوریتم‌های

یادگیری تقویتی کارآمدتر در محیط‌های پویا هموار می‌سازد [۲۰]. گائو و همکاران مسئله مسیریابی چند عاملی با رعایت محدودیت‌های زمانی را بررسی کردند و دو الگوریتم جستجوی مبتنی بر بهینه‌سازی و تقریب را ارائه دادند. نتایج نشان داد الگوریتم فوق تقریباً بهینه عمل کرده و در محیط‌های پر مانع (ثابت) موفقیت بالاتری دارد[۱۹]. ونو و گوروسامی یک مرور جامع بر الگوریتم‌های مسیریابی برای هدایت خودران ارائه دادند که روش‌های کلاسیک، فرا ابتکاری و مبتنی بر هوش مصنوعی را بررسی می‌کند. این مطالعه چالش‌های محیط‌های دینامیک، محدودیت‌های غیرهولونومیک[۷] و سطوح متفاوت دانش محیطی را تحلیل کرده و نقاط قوت و محدودیت هر دسته الگوریتم را برجسته می‌کند. همچنین، روندهای نوین مانند ادغام یادگیری ماشین و یادگیری تقویتی و مسیرهای تحقیقاتی آینده برای بهبود تطبیق‌پذیری و عملکرد سیستم‌های مسیریابی در محیط‌های پیچیده مورد بحث قرار گرفته است[۲۰]. آسلان و دمیرجی یک الگوریتم با یادگیری Q برای مسیریابی ربات ها ارائه کردند که علاوه بر بهینه‌سازی مسیر، مدیریت بیماری های همه گیر مانند قرنطینه و بازگشایی را نیز مد نظر دارد. این الگوریتم با بهبود روند درمان مدل شده و حذف پارامترهای کنترل اختصاصی، کارایی جستجو را افزایش می‌دهد. آزمایش‌ها در دوازده سناریوی مختلف نشان داد مسیرهای یافت‌شده توسط Q-LIPAعملکرد بهتری نسبت به سایر تکنیک‌های بهینه‌سازی هوشمند دارند و مدیریت مبتنی بر یادگیری تقویتی تأثیر مثبت بر عملکرد کلی الگوریتم دارد[۲۱]. وانگ و همکاران الگوریتم نوینی را برای مسیریابی ربات در محیط‌های پیچیده ارائه کردند که با نمونه برداری احتمالی منطقه‌ای، تنظیم گام تطبیقی، گرایش دینامیک به هدف و میدان پتانسیل مصنوعی بهینه‌شده، کارایی و همگرایی مسیر را بهبود می‌بخشد. این الگوریتم با استفاده از توسعه دو طرفه درخت بهینه‌سازی حریصانه و محدودیت‌های قابلیت اجرا، کیفیت مسیر و سرعت جستجو را افزایش می‌دهد. شبیه‌سازی‌ها نشان داد الگوریتم فوق نسبت به الگوریتم‌های معیار، مسیرهای کوتاه‌تر و زمان برنامه‌ریزی کمتری ارائه می‌دهد و برای هدایت ربات در محیط‌های پیچیده بسیار مناسب است[۲۲]. نی و همکاران یک الگوریتم ABC یکپارچه چند استراتژی مبتنی بر یادگیری Q (QMABC)برای مسیریابی وسایل نقلیه ارائه کردند که با بهره‌گیری از استراتژی‌های جستجوی متنوع و انتخاب هوشمندانه آن‌ها از طریق یادگیری Q، عملکرد جستجو را بهبود می‌بخشد. این الگوریتم با پیکربندی‌های مؤثرتر حالت[۸] و عمل، نسبت به روش‌های قبلی توانایی بهینه‌سازی بالاتری دارد. ارزیابی‌ها روی توابع معیار و کاربردهای مسیریابی نشان داد QMABC در حل مسائل عددی و عملی، عملکرد برتری نسبت به سایر الگوریتم‌های فراابتکاری دارد[۲۳]. محمد و همکاران یک کنترل‌کننده هوشمند برای موتور جت بر پایه یادگیری تقویتی عمیق و با استفاده از روش گرادیان قطعی سیاست عمیق ارائه کردند. نتایج شبیه‌سازی نشان داد این روش نسبت به کنترل‌کننده تناسبی–انتگرالی، زمان پاسخ کوتاه تر، کاهش دمای ورودی توربین و بهبود ایمنی عملکرد موتور را به همراه دارد. این پژوهش کارایی روش های یادگیری تقویتی در کنترل سامانه‌های غیرخطی و پیچیده را تأیید کرد[۲۴].

با وجود پژوهش‌های متعدد، مقایسه عملکرد یادگیری Q و یادگیری Q دوگانه در محیط‌های گروهی با موانع هر دو نوع ثابت و متحرک، همراه با ارزیابی قابلیت اطمینان مبتنی بر شبیه سازی مونت‌کارلو، مورد پژوهش قرار نگرفته است. همچنین، تأثیر افزودن درجات آزادی حرکتی (مانند حرکت مورب) بر کارایی و همگرایی این الگوریتم‌ها نیز نیاز به بررسی دارد. بنابراین، هدف اصلی این پژوهش، ارائه یک چارچوب ارزیابی جامع و مبتنی بر شبیه سازی مونت‌کارلو برای سنجش عملکرد و پایداری الگوریتم‌های یادگیری Q و یادگیری Q دوگانه در مسئله مسیریابی بلادرنگ گروهی ربات‌ها، در مواجهه با موانع ثابت و متحرک است.

ساختار این مقاله به این شرح است که در بخش دوم، مبانی نظری روش‌های یادگیری تقویتی مورد بحث قرار می‌گیرد. بخش سوم به تشریح دقیق محیط شبیه‌سازی، معماری حرکتی ربات‌ها، طراحی سازوکار پاداش و روش‌شناسی ارزیابی مونت‌کارلو می‌پردازد. در بخش چهارم، نتایج شبیه‌سازی و تحلیل‌های آماری ارائه می‌شود. بخش پنجم به بحث درباره یافته‌ها، دلالت‌های آن‌ها و مقایسه با کارهای مرتبط اختصاص دارد. در نهایت، در بخش ششم نتیجه‌گیری ارائه می‌شود.

۲. یادگیری تقویتی

۱,۲ یادگیری Q

یادگیری تقویتی[۹] یک چارچوب ریاضیاتی است که برای توسعه عامل‌های محاسباتی طراحی شده و امکان یادگیری رفتار بهینه را با ارتباط دادن پاداش‌های[۱۰] گذشته و جاری با اعمال انجام شده فراهم می‌کند. این چارچوب هوش مصنوعی در حوزه‌هایی مختلف علوم و مهندسی کاربردهای موفق فراوانی داشته و برای تصمیم‌گیری‌های پی‌درپی در محیط‌های ناشناخته و با داده‌های حجیم و بلادرنگ بسیار مناسب است. در این چارچوب، یادگیری Q به عنوان یک الگوریتم یادگیری تقویتی بدون مدل مطرح است که با استفاده از تابع Q، پاداش‌های احتمالی آینده اعمال در حالت‌های مشخص را تخمین می‌زند و از طریق به روز رسانی‌های تکراری، یادگیری سیاست بهینه را ممکن می‌سازد. الگوریتم یادگیری Q بر اساس معادله بلمن[۱۱] عمل می‌کند:

$$Q(s,a) = E[r + \gamma . max\, Q(s',a')] \quad (۱)$$

در معادله (۱)، s نشان دهنده حالت فعلی است و a عمل انتخاب شده در این حالت می‌باشد. پس از اجرای عمل a ، پاداش دریافتی با نماد r مشخص می‌شود و حالت بعدی که به آن منتقل می‌شود، با 's نمایش داده می‌شود. ضریب تنزیل γ تعادل بین پاداش‌های فوری و پاداش‌های آینده را تعیین می‌کند و $max\, Q(s',a')$ بهترین ارزش قابل دستیابی از حالت بعدی 's را نشان می‌دهد که بیشترین پاداش بلند مدت ممکن را منعکس می‌کند. این ضریب در بازه [۱ ۰] تعیین می‌کند که پاداش‌های آینده نسبت به پاداش فوری r چه وزنی دارند. عبارت $max\, Q(s',a')$ نیز بیانگر بیشترین ارزش قابل دستیابی از حالت بعدی الگوریتم با هدف یادگیری تابع $Q(s,a)$ یا تابع ارزش عمل–حالت[۱۲] کار می‌کند. این تابع، پاداش تجمعی و تعدیل شده آینده‌ای را که عامل از انجام عمل a در حالت s انتظار دارد، تخمین می‌زند. هدف نهایی یادگیری، تقریب تابع ارزش بهینه $Q(s,a)$ است که در معادله بلمن بیان

می‌شود. یادگیری این تابع ارزش از طریق یک فرآیند تکراری و مبتنی بر تجربه صورت می‌پذیرد. در هر گام از تعامل با محیط، پس از دریافت پاداش و مشاهده حالت جدید، مقدار $Q(s,a)$مربوطه با استفاده از قاعده به‌روز رسانی اختلاف زمانی[۱۳] اصلاح می‌شود.

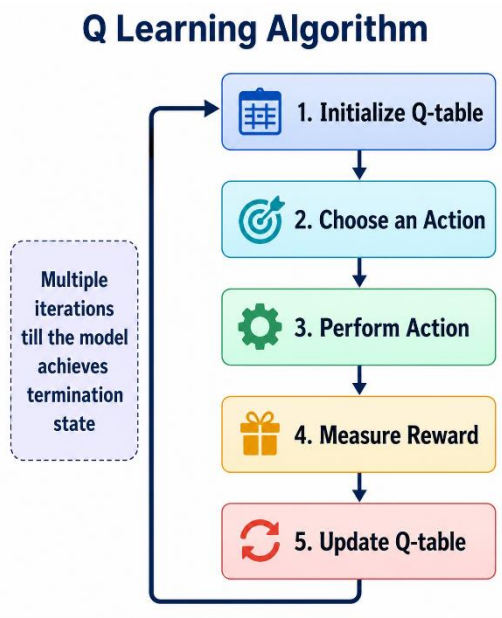
اجزای اصلی یادگیری Q شامل چند بخش مهم هستند. اولین بخش، مقادیر Q یا مقادیر عمل است که نشان‌دهنده پاداش‌های مورد انتظار برای انجام یک عمل در یک حالت مشخص می‌باشند. این مقادیر به مرور زمان و بر اساس قانون به‌روز رسانی اختلاف زمانی[۱۴] اصلاح می‌شوند. دوم، پاداش‌ها و اپیزودها[۱۵] هستند. در این بخش عامل با انجام اعمال مختلف از حالتی به حالت دیگر حرکت کرده و در این مسیر پاداش دریافت می‌کند. سوم، اختلاف زمانی به روز[۱۶] است که عامل برای به‌روز رسانی مقادیر Q از فرمول اختصاصی به‌روز رسانی اختلاف زمانی استفاده می‌کند، فرمولی که ترکیبی از پاداش فوری و تخمین پاداش‌های آینده را در نظر می‌گیرد.

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma max_a Q(S',A') - Q(S,A)] \quad (۲)$$

مجددا با باز تعریف فرمول (۱) و بیان سیاست نهایی یادگیری Q در فرمول (۲)، S نشان‌دهنده وضعیت فعلی عامل است و A عملی است که عامل در حالت S انجام می‌دهد. پس از انجام عمل، عامل به حالت بعدی S′ منتقل می‌شود و A′ بهترین عملی است که در حالت S′ می‌تواند انتخاب شود. R پاداشی است که عامل در نتیجه انجام عمل A در حالت S دریافت می‌کند و α نرخ یادگیری است که میزان تأثیر اطلاعات جدید بر مقادیر Q قدیمی را تعیین می‌کند. در الگوریتم های یادگیری تقویتی، سیاست (ε) برای ایجاد تعادل بین دانش فعلی و جست وجوی گزینه‌های جدید استفاده می‌شود. در این سیاست، عامل میان بهره‌برداری از دانش فعلی و اکتشاف گزینه‌های جدید تعادل برقرار می‌کند. عامل با احتمال $1-\varepsilon$ عملی را بر می‌گزیند که بیشترین مقدار Q را دارد و به این ترتیب پاداش مورد انتظار را حداکثر می‌کند. با این حال، با احتمال ε عمل به انتخاب تصادفی می‌کند تا مسیرهای ناشناخته را بررسی کرده و امکان دستیابی به راهبردهای بهتر را افزایش دهد. مدل‌های یادگیری Q یک فرایند تکراری را دنبال می کنند که در آن اجزای مختلف با هم همکاری می‌کنند تا عامل آموزش ببیند. به عبارت دیگر، عبارت داخل کروشه خطای تخمین یا اختلاف زمانی نامیده می‌شود که تفاوت بین تخمین جدید $[R + \gamma max_a Q(S',A') - Q(S,A)]$ و تخمین قدیمی $Q(S,A)$ است. الگوریتم با کاهش تدریجی این خطا، مقادیر Q را به سمت مقادیر بهینه همگرا می‌کند. انتخاب عمل در هر حالت، نقش تعیین کننده‌ای در فرآیند یادگیری دارد. برای ایجاد تعادل ضروری بین اکتشاف[۱۷] (جست‌وجوی فضای عمل) و بهره‌برداری[۱۸] (استفاده از دانش فعلی)، از سیاست ε حریصانه استفاده می‌شود. بر اساس این سیاست:

- با احتمال $1-\varepsilon$ عامل، عمل دارای بیشترین مقدار Q در حالت فعلی را انتخاب می‌کند (بهره‌برداری).
- با احتمال ε، عامل یک عمل را به‌طور کاملاً تصادفی از بین اعمال ممکن بر می‌گزیند (اکتشاف).

مقدار ε معمولاً در ابتدای یادگیری بالا است تا فضای حالت-عمل به خوبی کاوش شود و به تدریج کاهش می‌یابد تا عامل در مراحل پایانی بر دانش به دست آمده خود تکیه کند. این فرآیند تلفیق سیاست انتخاب عمل و قاعده به‌روز رسانی مقادیرQ ، هسته اصلی الگوریتم یادگیری Q را تشکیل می‌دهد. در ادامه، این روند به صورت مرحله به مرحله در شکل (۱) توضیح داده می‌شود.



شکل ۱. الگوریتم یادگیری Q

فرآیند یادگیری Q با شروع از یک حالت اولیه آغاز می‌شود که وضعیت محیط یا شرایط کنونی عامل را توصیف می‌کند. سپس عامل بر اساس حالت فعلی و سیاست خود یک عمل انتخاب می‌کند. این انتخاب معمولاً با استفاده از جدول Q انجام می‌شود که پاداش‌های احتمالی برای جفت‌های حالت-عمل را تخمین می‌زند و اغلب از استراتژی ε بهره می‌برد؛ به این صورت که با احتمال ε اعمال جدید را به‌طور تصادفی امتحان می‌کند تا مسیرها و راهبردهای تازه را کشف کند و با احتمال $1-\varepsilon$ از اعمالی استفاده می‌کند که تاکنون بیشترین پاداش را داشته‌اند، فرایندی که به آن بهره‌برداری گفته می‌شود. پس از انتخاب عمل، عامل آن را اجرا کرده و محیط به دو صورت واکنش نشان می‌دهد: ارائه یک حالت جدید S' به عنوان نتیجه عمل و ارائه یک پاداش R به عنوان بازخورد از کارایی عمل. با استفاده از این تجربه جدید، عامل جدول Q را به‌روز رسانی می‌کند. مقدار مربوط به جفت حالت-عمل براساس پاداش دریافتی و حالت جدید تنظیم می‌شود. این امر به عامل کمک می‌کند تخمین دقیق‌تری از سودمندی اعمال مختلف داشته باشد. با به‌روز رسانی مقادیرQ ، عامل سیاست خود را اصلاح کرده و چرخه مشاهده حالت‌ها، انتخاب اعمال، دریافت پاداش و به‌روز رسانی مقادیر Q را بارها تکرار می‌کند تا در نهایت سیاست بهینه‌ای را یاد بگیرد که بالاترین پاداش ممکن را در محیط به‌طور مداوم به دست آورد.

۲,۲ یادگیری Q دوگانه[19]

یادگیری Q دوگانه یک روش بهبود یافته از الگوریتم استاندارد یادگیری Q است که توسط هسلت در سال ۲۰۱۰ معرفی شد و برای حل مشکل بیش برآوردی[20] در مقادیر Q طراحی شده است. در یادگیری Q استاندارد، از همان مقادیرQ هم برای انتخاب عمل بهینه و هم برای ارزیابی ارزش آن عمل استفاده می‌شود که این مسئله می‌تواند منجر به بیش برآوردی در مقادیر Q گردد و در نتیجه یادگیری سیاست‌های زیر بهینه و کاهش پایداری الگوریتم را به دنبال داشته باشد. در فرمول به‌روز رسانی یادگیری Q یعنی معادله (۳) مقدار $max\ Q(s',a')$ هم عمل بهینه را انتخاب و هم ارزش آن را ارزیابی می‌کند. از آنجایی که مقادیر Q برآوردی هستند و شامل نویز می‌باشند، ماکزیمم گرفتن از این مقادیر

منجر به بیش برآوردی می‌شود، این بیش برآوردی به خصوص در محیط‌های پیچیده و با نویز می‌تواند مشکل‌ساز شود. ایده اصلی یادگیری دوگانه Qاستفاده از دو تابع ارزش مستقل (QAو QB) است. این دو تابع به صورت متناوب برای انتخاب عمل بهینه و ارزیابی ارزش آن استفاده می‌شوند. فرمول‌های به روز رسانی یادگیری Q دوگانه به این صورت می‌باشند:

$$E[max\ Q(s', a')] \geq max\ E[Q(s', a')] \quad (۳)$$

این نابرابری ریاضی توضیح می‌دهد که یادگیری Q معمولی می‌تواند مقدار Q را بیش‌برآورد کند. به عبارتی می توان گفت، انتظار از ماکزیمم، از ماکزیمم انتظارها بزرگ‌تر است.

$$a^* = argmax\ QA(s', a') \quad (۴)$$

a^* عمل بهینه در حالت بعدی s′ بر اساس یکی از دو شبکه Q است. به عبارتی انتخاب بهترین عمل در حالت بعدی بر اساس شبکه QA است.

$$QA(s,a) \leftarrow QA(s,a) + \alpha\,[r + \gamma\, QB(s', a^*) - QA(s,a)] \quad (۵)$$

از طرف دیگر معادله (۵) به‌روزرسانیQA با استفاده از مقدار هدف از شبکه QB است.

$$a^* = argmax\ QB(s', a') \quad (۶)$$

معادله (۶) انتخاب بهترین عمل در حالت بعدی بر اساس شبکه QB است.

$$QB(s,a) \leftarrow QB(s,a) + \alpha\,[r + \gamma\, QA(s', a^*) - QB(s,a)] \quad (۷)$$

معادله (۷)، به‌روزرسانی QB با استفاده از مقدار هدف از شبکه است. در این روش به جای یک شبکه ارزش‌گذاری، از دو شبکه QA و QBبه‌طور متناوب استفاده می‌شود. هنگام به‌روز رسانیQA ، برای محاسبه مقدار هدف از شبکه QBاستفاده می‌شود و بالعکس. این جداسازی باعث می‌شود که انتخاب عمل بهینه (argmax) و ارزیابی ارزش آن عمل توسط یک شبکه واحد انجام نشود، که عامل اصلی بیش‌برآورد در یادگیری Q معمولی است. با به‌روز رسانی متناوب این دو شبکه، یادگیری پایدارتر شده و مقادیرQ با دقت بیشتری به مقادیر بهینه واقعی همگرا می‌شوند.

۳. روند شبیه‌سازی یادگیری تقویتی برای مسیریابی بلادرنگ گروه ربات‌های هوایی در محیط پویا

شبیه‌سازی در یک محیط شبکه‌ای گسسته با ابعاد ۲۰ در ۲۰ سلول طراحی شده است. این فضا متشکل از ۱۰ ربات هوایی خودران، ۱۲ مانع ثابت و ۲ مانع متحرک است. هدف نهایی، رساندن تمامی ربات‌ها از موقعیت‌های آغازین پراکنده به یک موقعیت هدف ثابت بود. در طراحی یادگیری Q دو معماری حرکتی مجزا معماری چهار-جهته و معماری شش-جهته مورد بررسی قرار گرفت. در معماری اول، مجموعه عمل‌های هر ربات شامل حرکت به سمت بالا، راست، پایین و چپ بود. در معماری دوم، برای افزایش انعطاف در مسیریابی، دو حرکت مورب (بالا-راست و پایین-راست) نیز به مجموعه عمل‌ها افزوده شد. این طراحی امکان بررسی تأثیر درجه آزادی حرکتی بیشتر بر کارایی و پیچیدگی یادگیری را فراهم می‌ساخت.

موانع ثابت به صورت سلول‌های غیرقابل عبور در نظر گرفته شدند که موقعیت آن‌ها به صورت تصادفی، با اولویت قرارگیری در مناطق میانی و دور از نقاط شروع و هدف، تولید گردید. موانع

متحرک به صورت عامل‌های پویا شبیه‌سازی شده اند که در هر گام زمانی با احتمال غالب در جهت تعیین‌شده قبلی خود (یکی از چهار جهت اصلی) یک سلول جابه‌جا می‌شوند و با احتمال کم، جهت حرکت خود را به صورت تصادفی تغییر می‌دهند. حرکت این موانع در مرزهای محیط محدود شده و از برخورد با موانع ثابت یا اشغال سلول هدف جلوگیری می‌شد.

معماری این الگوریتم، یک ساختار سه‌بعدی بر اساس یادگیری Q و جدول مربوطه آن دارد. ابعاد این جدول به صورت (تعداد حالت‌ها × تعداد عمل‌ها × تعداد ربات‌ها) تعریف شد. هر درایه در این جدول، مقداری عددی را ذخیره می‌کرد که نشان‌دهنده برآورد عامل از سودمندی بلند مدت انجام یک عمل خاص در یک حالت خاص برای یک ربات خاص است. مقادیر اولیه این جدول با اعداد کوچک تصادفی مقدار دهی می‌شوند تا زمینه اولیه‌ای برای فرآیند اکتشاف فراهم گردد (به بخش پیشین یادگیری Q و Q دوگانه مراجع شود.). برای اعمال سیاست جهت کاهش برآورد بیش از حد، مقادیر Q که ممکن است در الگوریتم یادگیری Q رخ دهد، الگوریتم یادگیری Q دوگانه نیز پیاده‌سازی شده است. در این روش، به جای یک جدول Q، از دو جدول Q مستقل QA و QB استفاده گردید. در هر مرحله به‌روز رسانی، به صورت تصادفی انتخاب می‌شود که کدام یک از این دو جدول، عمل بهینه‌یابی را انجام دهد و کدام یک مقدار هدف را ارائه کند. این مکانیزم دوگانه، منجر به تخمین محافظه کارانه‌تر و پایدارتر مقادیر Q می‌شود.

یک ساختار پاداش چندلایه و سازگارشونده طراحی شده است تا یادگیری را هدایت کند. این ساختار شامل جریمه گام برای تشویق به کارایی، پاداش پیشرفت برای تقویت حرکت به سمت هدف (بر اساس کاهش فاصله اقلیدسی)، جریمه بالا برای برخورد با موانع، جریمه کم‌تر برای برخورد بین ربات‌ها، پاداش کلان دستیابی به هدف و در نهایت یک پاداش تشویقی گروهی برای مواردی که تمام ربات‌ها موفق به رسیدن به مقصد است. یادگیری در قالب دوره‌های آموزش متوالی انجام می پذیرد. در هر دوره، ربات‌ها از موقعیت‌های اولیه تا رسیدن به هدف یا رسیدن به حداکثر تعداد گام مجاز، به تعامل با محیط ادامه می‌دهند. انتخاب عمل در هر گام بر اساس سیاست اپسیلون-حریصانه[21] صورت می‌گرفت. در این سیاست، با احتمال $1-\varepsilon$ ربات عملی را انتخاب می‌کند که در جدول Q مربوط به خود، بیشترین مقدار را برای حالت فعلی دارد و با احتمال ε، یک عمل را به‌طور کاملاً تصادفی برمی‌گزینند. مقدارε به تدریج از یک مقدار آغازین بالا به یک مقدار پایین نهایی کاهش می‌یابد.

برای اطمینان از مقاوم بودن الگوریتم و کاهش وابستگی آن‌ها به شرایط تصادفی اولیه (مانند مکان موانع و مقداردهی اولیه جدول Q)، از روش شبیه‌سازی مونت‌کارلو استفاده شده است. این روش شامل اجرای ۲۰۰ تکرار مستقل کامل از فرآیند یادگیری است. در هر تکرار، تمام مراحل شامل تولید موانع، مقدار دهی اولیه جداول Q و اجرای ۸۰۰ دوره آموزشی، به‌طور مجزا و تصادفی انجام می‌گیرد. معیارهای عملکرد (میانگین پاداش، نرخ موفقیت، تعداد برخوردها) در هر دوره برای هر تکرار ذخیره می‌شود. پس از اتمام تمامی تکرارهای مونت‌کارلو، داده‌های خروجی برای هر یک از چهار شرایط آزمایشی (ترکیب دو الگوریتم و دو معماری حرکتی) پردازش می شود. برای هر معیار و در هر شماره دوره، میانگین و انحراف

معیار در بین ۲۰۰ تکرار محاسبه می شود. این موضوع منجر به تولید منحنی‌های یادگیری هموار و قابل اطمینانی می شود که روند همگرایی و عملکرد نهایی هر روش را به وضوح نشان می‌داد. همچنین، داده‌های ۱۰۰ دوره پایانی به عنوان شاخص عملکرد پایدار و نهایی هر الگوریتم استخراج می شود.

در مرحله نهایی، بر روی داده‌های عملکرد پایدار (دوره‌های پایانی) تحلیل‌های آماری پیشرفته انجام می گیرد. این تحلیل‌ها شامل مقایسه میانگین‌ها با استفاده از آزمون t مستقل، محاسبه اندازه اثر (به عنوان مثال، d کوهن) برای بررسی کیفی میزان برتری یک روش نسبت به دیگری و تحلیل نرخ همگرایی (شمارش تعداد تکرارهایی که به آستانه عملکرد مشخصی دست یافته‌اند) است. نتایج این تحلیل‌های کمی، مبنای علمی برای مقایسه نهایی کارایی و قابلیت اطمینان الگوریتم‌های مورد مطالعه فراهم می‌آورد.

در این پژوهش، به منظور هدایت گروهی ربات‌ها در محیط پویا با موانع متحرک، یک ساختار پاداش تطبیقی طراحی شده است. مقادیر عددی تمام پارامترهای ثابت پاداش (از جمله پاداش هدف، جریمه برخورد، جریمه گام و غیره) در جدول (۲) ارائه شده است. در ادامه، فرمول‌های ریاضی مربوط به اجزای پویای پاداش و مکانیزم‌های تطبیقی‌سازی نرخ یادگیری و نرخ کاوش ارائه می‌گردد. منظور از «پاداش تطبیقی»، تغییر پویای پارامترهای نرخ یادگیری و نرخ کاوش بر اساس پیشرفت فرآیند یادگیری است، نه تغییر خود پاداش. این مکانیزم باعث بهبود همگرایی و کاهش نوسانات در مراحل پایانی یادگیری می‌شود.

معادله (۸) پاداش پیشرفت بر اساس کاهش فاصله ربات تا هدف را تعریف می‌کند.

$$R_{progress} = 0.5 \times max(0, d_{current} - d_{next}) \quad (۸)$$

که در آن d فاصله اقلیدسی ربات تا هدف است. معادله (۹) نحوه محاسبه پاداش کل هر ربات در هر گام را نشان می‌دهد.

$$Rtotal = -0.05 + Rprogress + \begin{cases} +200 & \text{رسیدن به هدف} \\ -10 & \text{برخورد با مانع} \\ -2 & \text{برخورد با ربات دیگر} \end{cases} \quad (9)$$

همچنین معادله‌های (۱۰) و (۱۱) به ترتیب نرخ یادگیری تطبیقی و اپسیلون تطبیقی را بر اساس شماره اپیزود و عملکرد اخیر سیستم بیان می‌کنند:

$$\alpha_{adaptive} = 0.1 \times \left(1 - 0.5 \times \frac{ep}{800}\right) \quad (۱۰)$$

(۱۱)

$$\epsilon adaptive = \epsilon \times \left(1 - 0.3 \times (\text{میانگین موفقیت ۱۰ اپیزود اخیر})\right)$$

۴. آزمون‌های آماری مورد استفاده در تحلیل نتایج شبیه‌سازی

۱،۴ تحلیل توصیفی و محاسبه معیارهای مرکزی و پراکندگی

پیش از هر گونه آزمون استنباطی، برای هر یک از چهار گروه آزمایشی حاصل از ترکیب دو الگوریتم یادگیری Q و یادگیری Q دوگانه با دو معماری حرکتی ۴ و ۶ جهته، معیارهای آماری توصیفی محاسبه می‌گردد. این محاسبات بر روی داده‌های عملکرد نهایی (متوسط ۱۰۰ دوره آخر) تمامی اجراهای مونت‌کارلو انجام می شود. مهم‌ترین این معیارها شامل میانگین حسابی:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (۱۲)$$

به عنوان شاخص مرکزی و انحراف معیار:

$$= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad (۱۳)$$

به عنوان شاخص پراکندگی است. این مقادیر پایه لازم برای کلیه تحلیل‌های مقایسه‌ای بعدی را فراهم می نماید.

۲,۴ آزمون t مستقل برای مقایسه میانگین‌های دو گروه

برای بررسی وجود تفاوت آماری معنی‌دار بین میانگین عملکرد دو الگوریتم در هر معماری حرکتی، از آزمون t مستقل دو نمونه‌ای استفاده شده است.

این آزمون فرض صفر مبنی بر برابری میانگین‌های جامعه $H_0: \mu_1 = \mu_2$ را می‌آزماید. آماره t با فرض برابری واریانس‌ها (با استفاده از واریانس ترکیبی) به صورت زیر محاسبه می گردد.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (۱۴)$$

که در آن

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}} \quad (۱۵)$$

$\bar{X}_1 \bar{X}_2$ و بیان کننده میانگین‌های نمونه هستند.

این نمادها نشان‌دهنده میانگین حسابی (مقدار متوسط) داده‌های عملکرد در هر یک از دو گروه مستقل مورد مقایسه هستند. این میانگین‌ها از تقسیم مجموع مقادیر مشاهده‌شده در هر گروه بر تعداد اعضای آن گروه محاسبه می‌شوند و شاخص اصلی مرکزی ترین نقطه توزیع داده‌های آن گروه را نشان می‌دهند.

n_1 و n_2 نشان‌دهنده تعداد مشاهدات یا نمونه‌های مستقل در هر یک از دو گروه هستند. در چارچوب این تحقیق n_1 تعداد اجراهای مونت‌کارلو (تکرارهای مستقل شبیه‌سازی) در گروه اول است. در این راستا، n_2 تعداد اجراهای مونت‌کارلو در گروه دوم است. سپس مقدار احتمال p با درجه آزادی $df = n_1 + n_2 - 2$ از توزیع t استودنت استخراج شد. سطح معنی‌داری $\alpha = 0.05$ در نظر گرفته شد.

پس از محاسبه آماره آزمون t از روی مقادیر میانگین‌ها، انحراف معیارها و حجم نمونه‌های دو گروه، گام نهایی استنباط آماری با تعیین درجه آزادی $df = n_1 + n_2 - 2$ و استخراج مقدار احتمال متناظر p از تابع چگالی احتمال توزیع t استودنت با همان درجه آزادی انجام می‌پذیرد. این مقدار p که تحت فرض صفر برابری میانگین‌های جامعه محاسبه می‌شود، بیانگر احتمال نظری مشاهده‌ی تفاوتی به اندازه‌ی تفاوت نمونه‌ای مشاهده شده یا صرفاً بر اثر نوسانات تصادفی نمونه‌گیری است. برای اتخاذ تصمیم نهایی، این مقدار p با یک آستانه از پیش تعیین شده به نام سطح معناداری α که در این پژوهش مطابق استاندارد رایج علوم مهندسی برابر ۰,۰۵ (۵٪) در نظر گرفته شده است، مقایسه می‌گردد. چنانچه رابطه $p < \alpha$ برقرار باشد، دال بر آن است که احتمال رخداد تصادفی تفاوت مشاهده شده به اندازه‌ی کافی پایین بوده و بنابراین فرض صفر مبنی بر عدم تفاوت معنادار بین میانگین‌های دو گروه مردود اعلام می‌شود و نتیجه‌گیری به نفع وجود تفاوت آماری معنادار می‌گردد.

۳,۴ محاسبه انـدازه اثـر d کـوهن[۲۲] بـرای کمّی‌سازی بزرگی تفاوت

از آنجا که آزمون t تنها معنی‌داری آماری را نشان می‌دهـد، بـرای ارزیـابی اهمیـت عملـی و بزرگـی تفاوت بین دو گروه، اندازه اثـر اسـتاندارد شـده d کـوهن محاسـبه گردیـد. ایـن معیـار، تفـاوت میانگین‌ها را بر حسب واحد انحراف معیار ترکیبی بیان می‌کند:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \quad (۱۶)$$

مقـادیر d بـر اسـاس راهنمـای کـوهن بـه صـورت $|d| \approx 0.2$ اثـر کوچـک، $|d| \approx 0.5$ اثـر متوسط و $|d| \geq 0.8$ (اثر بزرگ) تفسیر می شوند. این محاسبه امکان مقایسه مستقیم میزان برتـری یک روش نسبت به دیگری را فراهم می آورد.

۴,۴ تحلیل نرخ همگرایـی بـا شـاخص‌هـای کسری

برای ارزیابی سرعت و قابلیت اطمینـان همگرایـی الگوریتم‌ها، از تحلیل‌های مبتنی بر نسبت استفاده مـی‌شـود. معیـار اصـلی، نسـبت اجراهـای مونت‌کارلو است که تا یک آستانه عملکـرد خـاص (مثلاً دستیابی بـه نـرخ موفقیـت بـالای ۵۰٪) در یک دوره زمـانی مشـخص (مـثلاً ۱۰۰ دوره اول) همگرا شده انـد. بـازه اطمینـان ۹۵٪ بـرای ایـن نسبت $\hat{p}$(نرخ همگرایی مشاهده شده) با اسـتفاده از:

$$\hat{p} \pm Z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (۱۷)$$

محاسبه می شود تا عـدم قطعیـت در بـرآورد قابلیت اطمینان هر روش نیز کمّی‌سازی گردد.

۵,۴ آزمون فرض برابری واریانس‌ها (آزمـون لوین[۲۳])

به عنوان پیش‌نیاز و بررسی فرض ضمنی آزمون t مبنی بر همگنی واریانس‌هـا، آزمـون لـوین بـرای مقایسه واریانس دو گروه اجرا می گردد. آماره این آزمون بر اساس انحـراف‌هـای مطلـق از میـانگین محاسبه و مقدار p مربوطه استخراج می شود. این رویکـرد، جامعیـت و اتکاپـذیری تحلیـل‌هـای مقایسـه‌ای را تضـمین مـی کنـد. در جـدول (۱) خلاصه این آزمون ها و نحـوه بررسـی آنهـا آورده شده است.

جدول۱. خلاصه آزمون‌های آماری مورد استفاده در تحلیل نتایج شبیه‌سازی

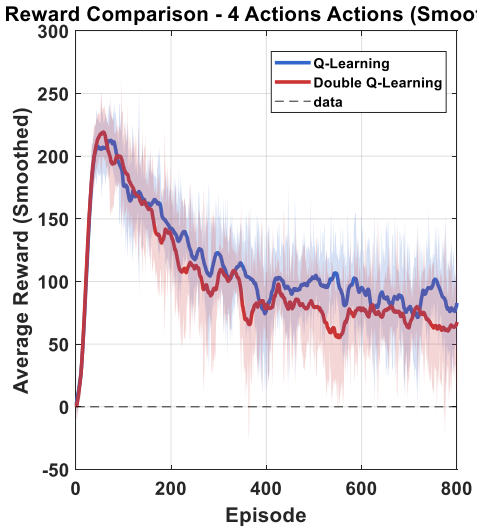
نام آزمون / شاخص	هدف از استفاده	فرمول محاسباتی	خروجی‌های اصلی	نحوه بررسی و تفسیر
معیارهای توصیفی	خلاصه‌سازی و توصیف داده‌های عملکرد هر گروه آزمایشی.	میانگین: $\bar{X} = \frac{1}{n}\sum X_i$ انحراف معیار: $s = \sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}$	میانگین و انحراف معیار برای هر گروه.	مقایسه مستقیم عددی میانگین‌ها و انحراف معیارها. گروه با میانگین بالاتر و انحراف معیار پایین‌تر، عملکرد بهتر و پایدارتری دارد.

تحلیل نرخ همگرایی	آزمون کولموگروف-اسمیرنوف (برابری واریانس)
ارزیابی قابلیت اطمینان و سرعت دستیابی به سطح عملکرد قابل قبول توسط الگوریتم‌ها.	بررسی فرض همگنی واریانس‌ها به عنوان پیش‌نیاز استفاده از نسخه استاندارد آزمون t .
فاصله اطمینان $\hat{p} \pm Z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ نرخ همگرایی: $\hat{p} = \frac{\text{تعداد اجراهای همگرا}}{\text{کل اجراها}}$	بر اساس مقیاس انحراف‌های مطلق از میانگین یا میانه.
نسبت $\hat{p}$ و فاصله اطمینان ۹۵٪ حد پایین و حد بالا	آماره آزمون و مقدار p-value مربوطه
مقایسه مستقیم نرخ‌های $\hat{p}$ روش با نرخ بالاتر، قابل اطمینان‌تر است. عدم همپوشانی فاصله اطمینان دو گروه نشان‌دهنده تفاوت معنی‌دار در قابلیت اطمینان آن‌ها است.	اگر $p\text{-value} < 0.05$، فرض برابری واریانس‌ها رد می‌شود. در این صورت باید از نسخه اصلاح‌شده آزمون t مانند آزمون ولچ با درجه آزادی تعدیل‌شده استفاده کرد.

آزمون t مستقل	اندازه اثر (کوهن d)
بررسی وجود تفاوت آماری معنی‌دار بین میانگین‌های دو گروه مستقل (QL در مقابل DQL)	$t = \frac{}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ $s_p = \sqrt{\frac{}{n_1+n_2-2}}$
آماره t، درجه آزادی df، و مقدار p-value	$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}$
اگر $p\text{-value} < 0.05$ تفاوت از نظر آماری معنی‌دار است. مقدار t علامت جهت تفاوت را نشان می‌دهد (مثبت بودن به نفع گروه اول).	مقدار عددی d می‌تواند مثبت یا منفی باشد [illegible] $\lvert d \rvert \approx 0.2$: اثر کوچک $\lvert d \rvert \approx 0.5$: اثر متوسط $\lvert d \rvert \geq 0.8$: اثر بزرگ هر چه قدر مطلق d بزرگ‌تر باشد، تفاوت عملی مهم‌تر است.

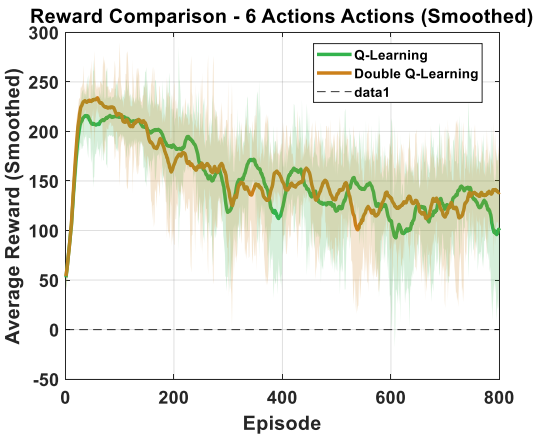
۵. بررسی نتایج

در این بخش، نمودارهای حاصل از شبیه‌سازی‌ها به تفصیل ارائه و تحلیل می‌شوند.



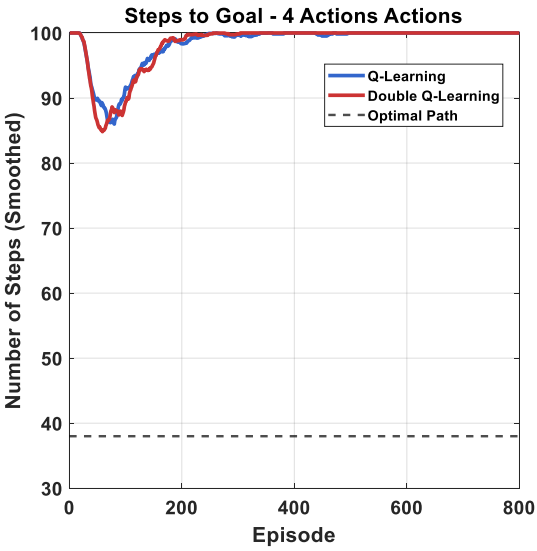
شکل ۲. مقایسه پاداش میانگین در حالت ٤ عمل

در شکل (۲)، نمودار روند پاداش میانگین در طول ۸۰۰ اپیزود یادگیری برای دو الگوریتم یادگیری Q و یادگیری Q دوگانه در حالت ۴ عمل نمایش داده شده است. همان‌طور که مشاهده می‌شود، منحنی مربوط به یادگیری Q (خط آبی) با فاصله محسوسی از منحنی یادگیری Q دوگانه (خط قرمز) قرار دارد و رفتار هردو الگوریتم نوسانی است. پاداش نهایی یادگیری Q و یادگیری دوگانه Q تقریبا برابر است. این نتیجه حاکی از آن است که در یک فضای عمل بر اساس چهار حرکت اصلی، الگوریتم یادگیری Q و یادگیری Q دوگانه فرق چندانی باهم ندارند. در شکل (۲) میزان نهایی پاداش برای روش های یادگیری Q و یادگیری دوگانه Q به ترتیب برابر ۷۳,۱ ± ۹۴,۶ و ۷۳,۹ ± ۸۱,۹ است.



شکل ۳. مقایسه پاداش میانگین در حالت ۶ عمل

در شکل (۳)، روند پاداش در حالت ۶ عمل برای دو الگوریتم نشان داده شده است. همان طور که مشاهده می‌شود، عملکرد هر دو الگوریتم بسیار نزدیک بهم می‌باشد که نشان‌دهنده کارایی مشابه آن‌ها در فضای عمل پیچیده‌تر است. با این حال، یادگیری Q دوگانه اندکی از یادگیری Q پیشی گرفته است و در انتها به ارزش عمل بیشتری دست یافته است. در شکل (۳) میزان نهایی پاداش برای روش های یادگیری Q و یادگیری دوگانه Q به ترتیب برابر۱۴۲,۶ و ۱۴۷,۴ است.

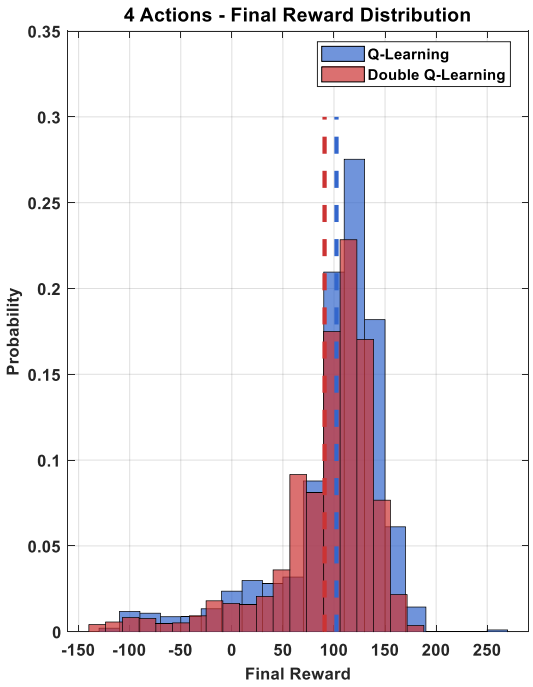


شکل ٤. تعداد گام‌ها تا هدف در حالت ٤ عمل

در شکل (۴)، روند کاهشی تعداد گام‌های لازم برای دستیابی به هدف در محیط ۴ عمل نمایش داده شده است. همان‌گونه که مشاهده می‌شود، هر دو الگوریتم پس از اپیزود ۳۰۰ به مرحله‌ای از پایداری و همگرایی رسیده‌اند. روند یادگیری این الگوریتم‌ها ابتدا نزولی، سپس صعودی و در نهایت پایدار است که نشان‌دهنده پیداکردن مسیر بهینه توسط این دو الگوریتم است.

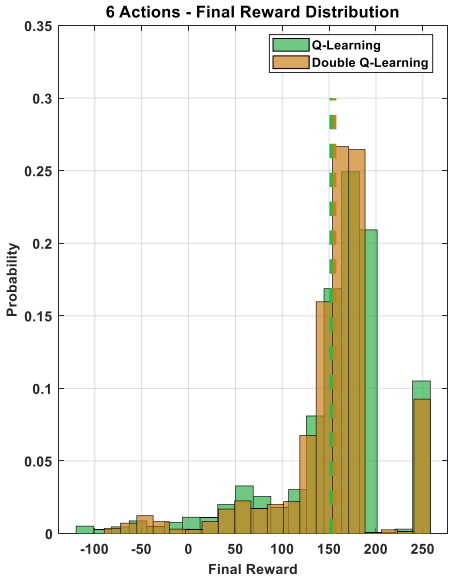
شکل۵. تعداد گام‌ها تا هدف در حالت ۶ عمل

در شکل (۵) تغییرات تعداد گام‌های لازم برای رسیدن به هدف در طول اپیزودها برای الگوریتم‌های یادگیری Q و یادگیری Q دوگانه با ۶ عمل نمایش داده شده است. در ابتدای فرآیند، هر دو الگوریتم با کاهش سریع تعداد گام‌ها، بهبود قابل‌توجهی در عملکرد نشان می‌دهند که بیانگر یادگیری اولیه مؤثر است. با افزایش اپیزودها، روند تغییرات به تدریج افزایشی و همراه با نوسان می‌شود که می‌تواند ناشی از اکتشاف مداوم و پیچیدگی فضای حالت–کنش باشد. در اغلب بازه‌ها، یادگیری Q و یادگیری Q دوگانه فاصله کمی با یکدیگر دارند. اما با این حال پس از ۴۰۰ اپیزود به روند همگرایی پیرامون ۹۰ رسیده‌اند.



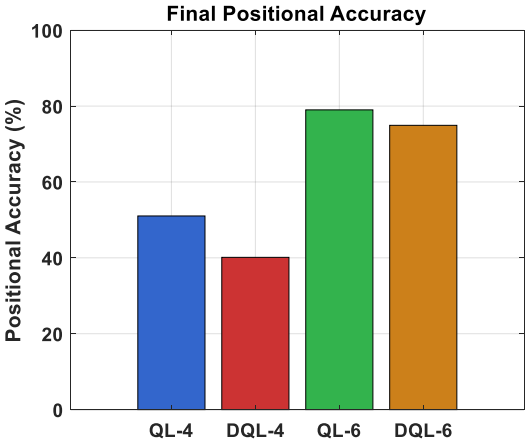
شکل۶. توزیع احتمال (نزدیک به نرمال) نهایی پاداش در حالت ٤ عمل

در شکل (۶)، توزیع احتمال نهایی پاداش برای دو الگوریتم در حالت ۴ عمل مقایسه شده است. این الگو تأیید می‌کند، یادگیری Q در بهترین حالت نیازمند به پاداش بالاتری بوده و نوسان بیشتری تجربه کند. در مقابل، یادگیری Q دوگانه عملکردی بهتر ارائه می‌دهد و پاداش نهایی آن کمتر از یادگیری Q است.



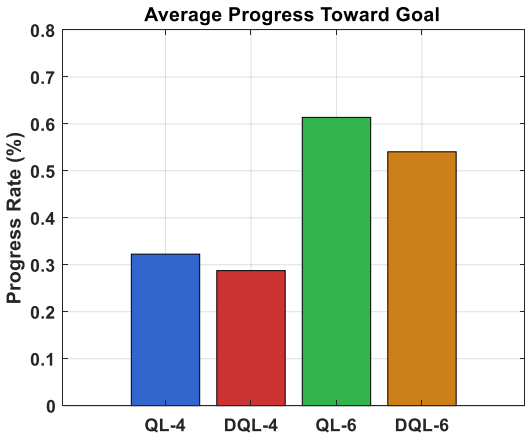
شکل۷. توزیع احتمال نهایی پاداش در حالت ۶ عمل

در شکل (۷) توزیع پاداش نهایی برای الگوریتم‌های یادگیری Q و یادگیری Q دوگانه در حالت ۶ عمل نمایش داده شده است. تمرکز اصلی هر دو توزیع در ناحیه پاداش‌های بالا قرار دارد که نشان‌دهنده دستیابی به سیاست‌های بهینه در انتهای فرآیند یادگیری است. با این حال، توزیع مربوط به یادگیری Q دوگانه فشرده‌تر و با پراکندگی کمتر مشاهده می‌شود که بیانگر پایداری بیشتر نتایج نهایی است. هم‌پوشانی قابل‌توجه دو توزیع نشان می‌دهد عملکرد میانگین هر دو الگوریتم نزدیک است.



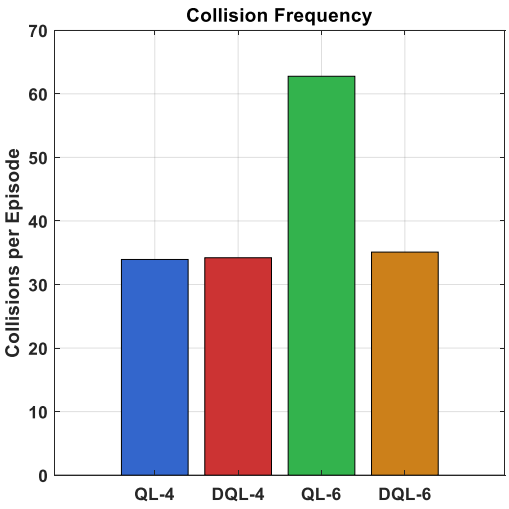
شکل۸. مقایسه دقت موقعیتی نهایی

در شکل (۸)، دقت موقعیتی نهایی (درصد موفقیت) چهار سناریوی شبیه‌سازی مقایسه شده است. همان طور که مشاهده می‌شود، عملکرد هر دو الگوریتم با افزایش پیچیدگی فضای عمل از ۴ به ۶ عمل بهبود چشمگیری یافته است. در حالت ۴ عمل، یادگیری Q از یادگیری Q دوگانه پیشی گرفته است. اما در محیط پیچیده‌تر ۶ عمل، این اختلاف تقریباً از بین رفته و هر دو الگوریتم به دقت‌های بسیار نزدیک و بالایی (حدود ۸۰٪) دست یافته‌اند. این نتیجه نشان می‌دهد که مزیت کاهش بیش‌برآوردی در یادگیری Q دوگانه در محیط‌های پیچیده‌تر خود را نمایان می‌سازد و باعث می‌شود کارایی آن به سطح الگوریتم استاندارد برسد.



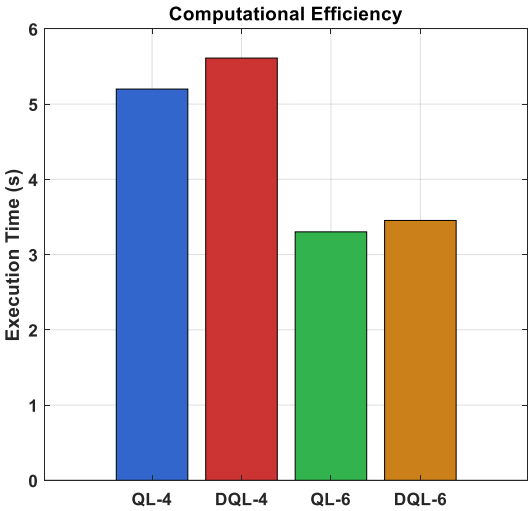
شکل۹. نرخ متوسط پیشرفت

در شکل (۹)، نرخ متوسط پیشرفت به سمت هدف برای چهار سناریو مقایسه شده است. همان‌طور که مشاهده می‌شود، الگوریتم‌های با ۶ عمل نرخ پیشرفت بسیار بالاتری (حدود ۰,۶٪) نسبت به الگوریتم‌های با ۴ عمل (حدود ۰,۳٪) دارند. این نشان می‌دهد فضای عمل غنی‌تر از بابت در اختیار داشتن حرکات مورب، امکان حرکت کارآمدتر به سمت هدف را فراهم می‌کند. در هر دو پیکربندی، عملکرد یادگیری Q و یادگیری Q دوگانه بسیار نزدیک به هم است، که حاکی از کارایی مشابه آن‌ها در بهینه‌سازی پیشرفت است. با این حال، در حالت ۶ عمل، یادگیری Q دوگانه اندکی بهتر عمل کرده که همسو با برتری آن در دقت موقعیتی است.



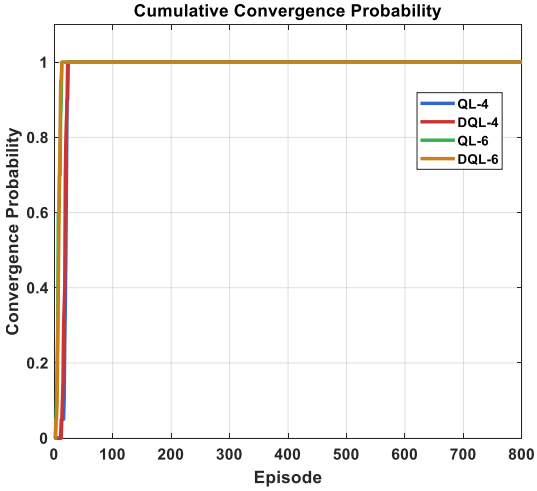
شکل ۱۰. مقایسه میانگین تعداد برخورد

در شکل (۱۰)، میانگین تعداد برخورد برای چهار سناریو مقایسه شده است. همان‌طور که مشاهده می‌شود، در هر دو پیکربندی، الگوریتم یادگیری Q دوگانه برخورد کمتری نسبت به یادگیری Q ایجاد کرده است. این بهبود ایمنی به ویژه در پیکربندی پیچیده‌تر ۶ عمل چشمگیرتر است. نکته مهم، افزایش کلی برخوردها در محیط ۶ عمل نسبت به ۴ عمل است که احتمالاً ناشی از فضای عمل بزرگتر و مسیرهای پیچیده‌تر است. با این وجود، یادگیری Q دوگانه در هر دو محیط، گزینه ایمن‌تری محسوب می‌شود.



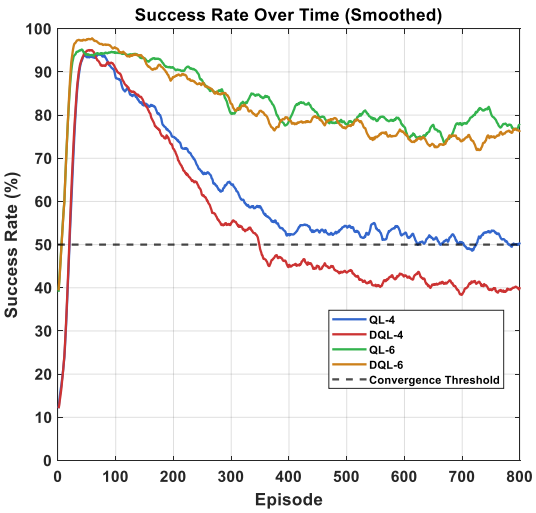
شکل ۱۱. کارایی محاسباتی

در شکل (۱۱)، زمان اجرای الگوریتم‌ها برای چهار سناریو مقایسه شده است. همان‌طور که مشاهده می‌شود، الگوریتم‌های با ۶ عمل به‌طور قابل توجهی سریع‌تر از الگوریتم‌های با ۴ عمل اجرا شده‌اند، که ناشی از همگرایی سریع‌تر در فضای عمل غنی‌تر است. در هر دو پیکربندی، زمان اجرای یادگیری Q دوگانه و یادگیری Q بسیار نزدیک به هم است، که نشان می‌دهد عملکرد Q دوگانه، هزینه محاسباتی اضافه‌ای تحمیل نمی‌کند. این نتیجه برای کاربردهای بلادرنگ بسیار مطلوب است.



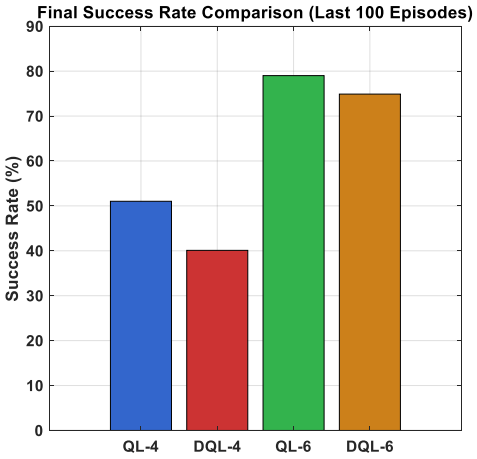
شکل۱۲. احتمال همگرایی تجمعی الگوریتم‌ها با آستانه ۵۰٪ موفقیت

در شکل (۱۲)، احتمال همگرایی تجمعی الگوریتم‌ها در طول اپیزودها نمایش داده شده است. همان‌طور که مشاهده می‌شود، منحنی‌های مربوط به معماری ۶ عمل (هر دو الگوریتم) شیب تندتری داشته و سریع‌تر به احتمال همگرایی ۱ می‌رسند. این نشان‌دهنده سرعت یادگیری بالاتر در فضای عمل است. با این حال، در نهایت هر چهار منحنی به همگرایی کامل می‌رسند که نشان‌دهنده کارایی کلی روش‌هاست.



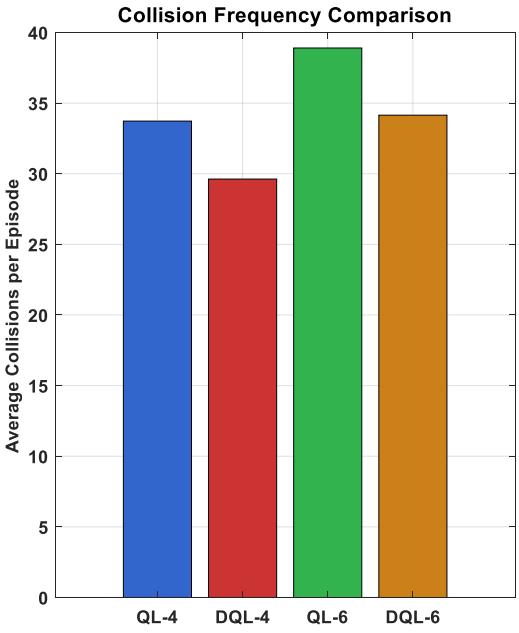
شکل۱۳. نرخ موفقیت بر حسب اپیزود

در شکل (۱۳)، روند نرخ موفقیت در طول فرآیند یادگیری برای چهار سناریو نمایش داده شده است. همان‌طور که مشاهده می‌شود، منحنی‌های مربوط به ۶ عمل از اپیزود ۱۰۰ به بعد بالاتر از منحنی‌های به ۴ عمل بوده و به نرخ موفقیت حدود ۸۰٪ رسیده‌اند. در مقابل، منحنی‌های ۴ عمل رشد کندتری داشته و در نهایت به نرخ موفقیت حدود ۵۰٪ الی ۶۰٪ محدود شده‌اند. در محیط ۴ عمل، یادگیری Q در تمام مراحل از یادگیری Q دوگانه پیشی گرفته است. این نتیجه به وضوح نشان می‌دهد که فضای عمل مهم‌ترین عامل در بهبود نرخ موفقیت نهایی سیستم است.



شکل۱٤. مقایسه نرخ موفقیت نهایی در ۱۰۰ اپیزود پایانی

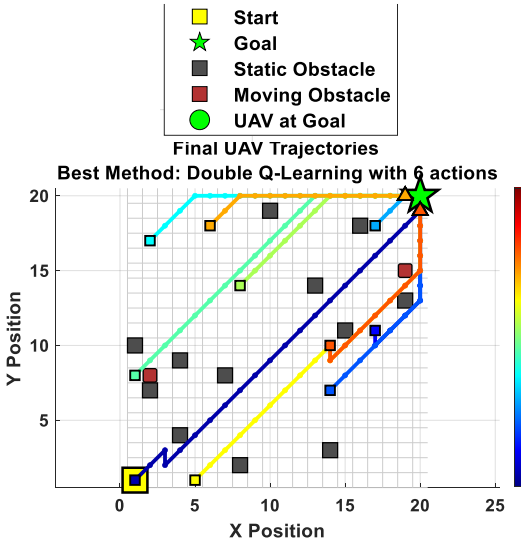
در شکل (۱۴)، نرخ موفقیت نهایی در مرحله پایانی یادگیری مقایسه شده است. همان طور که مشاهده می‌شود، عملکرد الگوریتم‌های ۶ عمل با اختلاف زیادی برتر از الگوریتم‌های ۴ عمل است. در حالت ۴-عمل، یادگیری Q حدود ۱۰٪ از یادگیری Q دوگانه پیشی گرفته است. اما در حالت ۶-عمل، این تفاوت تقریباً ناچیز شده و هر دو الگوریتم به نرخ موفقیت حدود ۸۰٪ دست یافته‌اند. این نتیجه تأکید می‌کند که در یک محیط عملیاتی پیچیده، یادگیری Q دوگانه می‌تواند بدون کاهش محسوس در نرخ موفقیت، به عنوان گزینه پایدارتر مورد استفاده قرار گیرد.



شکل۱۵. مقایسه میانگین برخورد

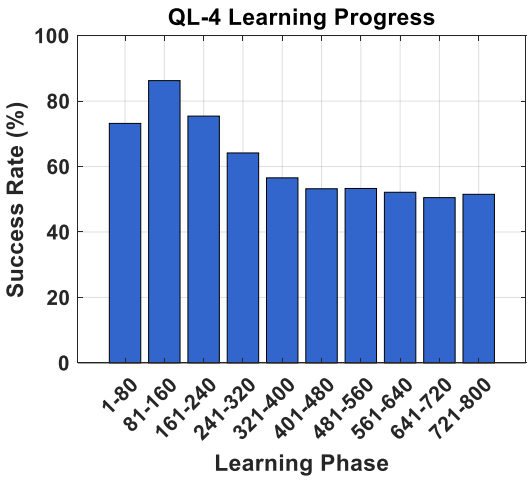
در شکل (۱۵)، میانگین تعداد برخورد در هر اپیزود برای چهار سناریو مقایسه شده است. همان‌طور که مشاهده می‌شود، در هر دو حالت ۴ و ۶ عمل، الگوریتم یادگیری Q دوگانه به‌طور محسوسی برخورد کمتری نسبت به یادگیری Q ثبت کرده است. این برتری در حالت پیچیده‌تر ۶ عمل بارزتر است. همچنین، محیط ۶ عمل علی‌رغم موفقیت بالاتر، میانگین برخورد بیشتری

نسبت به محیط ۴ عمل دارد که نشان‌دهنده تبادل بین پیچیدگی مسیر و ایمنی در این محیط است.



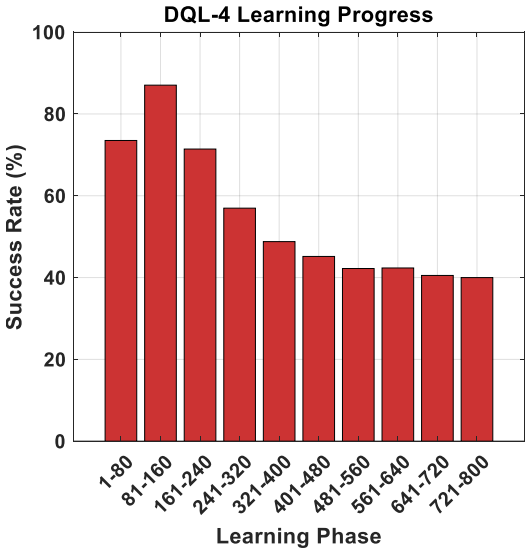
شکل۱۶. مسیرهای نهایی پیموده شده ربات‌ها

در شکل (۱۶) مسیرهای نهایی پیموده شده توسط رباتها تحت الگوریتم یادگیری Q دوگانه ترسیم شده است. همان‌طور که مشخص است، تمامی رباتها (که با اندیس‌های رنگی متفاوت تفکیک شده‌اند) توانسته‌اند با موفقیت و بدون برخورد با موانع ثابت و متحرک، خود را به نقطه هدف برسانند که این امر همگرایی کامل روش پیشنهادی را در محیط‌های پیچیده چندعاملی تأیید می‌کند.



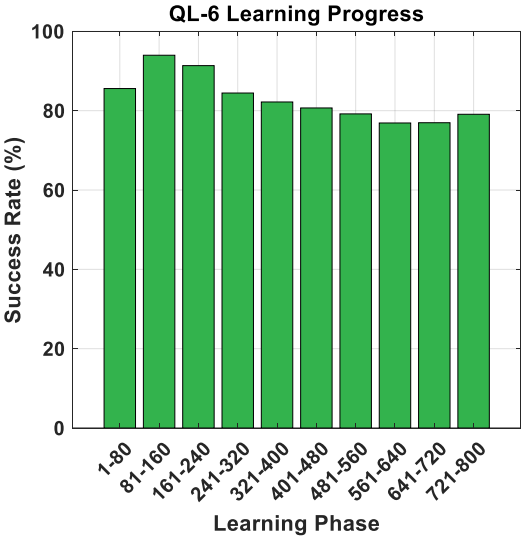
شکل ۱۷. نرخ موفقیت یادگیری Q در حالت ٤ عمل

در شکل (۱۷) روند پیشرفت یادگیری Q در یک محیط با چهار عمل ممکن نمایش داده شده است. همان‌گونه که مشاهده می‌شود. این الگو نشان می‌دهد که عامل به تدریج و از طریق تعامل با محیط، در حال یادگیری یک سیاست مؤثر برای بیشینه کردن پاداش تجمعی است. با این حال، نوسانات موجود در منحنی، خصوصاً در بازه‌های میانی، نشان‌دهنده‌ی ماهیت اکتشافی الگوریتم و فرآیند به‌روز رسانی ارزش‌های عمل-حالت باشد. در بازه ۸۱ تا ۱۶۱ عامل به بالاترین نرخ موفقیت دست یافته است.



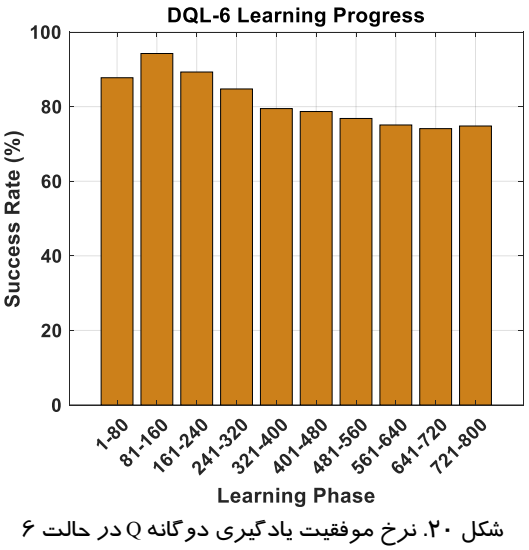
شکل ۱۸.نرخ موفقیت یادگیری Q دوگانه در حالت ٤ عمل

روند پیشرفت الگوریتم یادگیری دوگانه Q با ۴ عمل بر اساس نرخ موفقیت در فازهای مختلف در شکل (۱۸) نمایش داده شده است. داده‌ها نشان می‌دهند که سیستم پس از رسیدن به اوج عملکرد در فاز اولیه (بازه ۱۶۰-۸۱)، روندی نزولی را طی کرده و در نهایت نرخ موفقیت آن در حدود ۵۰ درصد تثبیت می‌شود که بیانگر محدودیت فضای کنش‌ها در حفظ پایداری و همگرایی طولانی‌مدت است.



شکل۱۹. نرخ موفقیت یادگیری Q در حالت ۶ عمل

در شکل (۱۹) عملکرد الگوریتم یادگیری Q با ۶ عمل در طول فازهای یادگیری به تصویر کشیده شده است. نتایج حاکی از آن است که سیستم پس از دستیابی به اوج کارایی در فازهای ابتدایی، توانسته است نرخ موفقیت خود را در سطحی مطلوب (حدود ۸۰ درصد) تثبیت کند که این امر نشان‌دهنده پایداری بهتر و تعادل مناسب‌تر فضای کنش‌ها نسبت به مدل‌های محدودتر است.



شکل ۲۰. نرخ موفقیت یادگیری دوگانه Q در حالت ۶ عمل

در شکل (۲۰) روند یادگیری الگوریتم یادگیری دوگانه Q با ۶ عمل نمایش داده شده است. این نمودار پایدارترین عملکرد را میان روش‌های بررسی شده نشان می‌دهد؛ به‌طوری‌که سیستم پس از اوج‌گیری اولیه، توانسته است نرخ موفقیت خود را بدون افت محسوس در سطحی بالا (بیش از ۸۰ درصد) حفظ کند که بیانگر تعادل عالی و کارایی برتر این الگوریتم در طولانی‌مدت است.

۶. مقایسه یادگیری Q و یادگیری Q دوگانه

در بخش‌های قبل، مبانی یادگیری Q و یادگیری Q دوگانه مورد بررسی قرار گرفت. در این بخش، به بررسی و مقایسه عملکرد دو الگوریتم یادگیری Q و یادگیری Q دوگانه در دو حالت مختلف پرداخته خواهد شد. هدف از این مقایسه، ارزیابی تأثیر تعداد و نوع حرکت‌ها بر کیفیت یادگیری، سرعت همگرایی و پایداری الگوریتم‌ها است. این سه حالت به گونه‌ای طراحی شده‌اند که پیچیدگی عمل ها[۲۴] را به تدریج افزایش دهند و امکان بررسی جامع رفتار الگوریتم‌ها را فراهم آورند.

جدول۲. خلاصه پارامترهای شبیه‌سازی و نتایج کلی

دسته	پارامتر	مقدار / نتیجه
پارامترهای محیط	اندازه محیط	۲۰ × ۲۰
	حالت شروع	۱
	حالت هدف	۴۰۰
	تعداد ربات ها	۱۰
	تعداد اپیزودهای یادگیری	۸۰۰
	تعداد اجراهای مونت کارلو	۲۰
	موانع ثابت	۱۲
	موانع متحرک	۲
	نرخ یادگیری	۰،۱
	ضریب تخفیف	۰،۹۵
	پاداش رسیدن به هدف	۲۰۰
	جریمه هر گام	−۰،۰۵
	جریمه برخورد با مانع	−۱۰
	جریمه برخورد ربات به ربات	−۲
	پاداش موفقیت	۵۰
	پاداش پیشرفت	۰،۵
نتایج	روش بهینه پیشنهادی	یادگیری Q

کلی و بهی‌نه		دوگانه با ۶ عمل
	دقت موقعی‌تی بهی‌نه	٪۱۰۰
	می‌انگی‌ن دقت موقعی‌تی کل	٪۶۵،۳۶
	بهبود ۶-عمل نسبت به ۴-عمل	٪۲۸،۰۷
	تعداد مراحل برای رسی‌دن همه ربات ها به هدف	۱۹ گام
تحلی‌ل همگرای‌ی	نرخ همگرای‌ی ۴-عمل (هر دو الگوری‌تم)	٪۱۰۰
	نرخ همگرای‌ی ۶-عمل (هر دو الگوری‌تم)	٪۱۰۰

جدول۳. نتایج تحلیل آماری مقایسه‌ای

پارامتر مقای‌سه	مقای‌سه ۴-عمل ی‌ادگی‌ری Q و ی‌ادگی‌ری Q دوگانه	مقای‌سه ۶-عمل ی‌ادگی‌ری Q و ی‌ادگی‌ری Q دوگانه
تفاوت معنی‌دار آماری(p)	۰،۰۰۰۰۰۰ بله	۰،۰۷۹۲۶۱ خی‌ر
مقدارt	۵،۴۴۶۴	۱،۳۲۵۶
اندازه اثر (Cohen's d)	۰،۱۷۲-	۰،۰۵۶
تفسی‌ر اندازه اثر	اثر ناچی‌ز	اثر ناچی‌ز
تفاوت در نرخ موفقی‌ت	ی‌ادگی‌ری Q بهتر است (+٪۸،۰۵)	ی‌ادگی‌ری Q دوگانه بهتر است (+٪۰،۴۳)
نتی‌جه گی‌ری عملکردی	ی‌ادگی‌ری Q دوگانه عملکرد ضعی‌ف‌تری دارد.	تفاوت معنی‌داری وجود ندارد. عملکرد ی‌ادگی‌ری Q دوگانه بهتر است.

جدول٤. مقایسه الگوریتم‌ها از نظر زمان آموزش، حافظه مصرفی و تعداد پارامترها

ردی‌ف	الگوری‌تم	تعداد پارامترها	حافظه مصرفی تقری‌بی (کی‌لوبای‌ت)	زمان آموزش (ثانی‌ه)
۱	ی‌ادگی‌ری Q (۴ حرکت)	۱۶۰۰۰	۱۲۸	۵،۲۷
۲	ی‌ادگی‌ری Q دوگانه (۴ حرکت)	۳۲۰۰۰	۲۵۶	۵،۶۹
۳	ی‌ادگی‌ری Q (۶ حرکت)	۲۴۰۰۰	۱۹۲	۳،۴۵
۴	ی‌ادگی‌ری Q دوگانه (۶ حرکت)	۴۸۰۰۰	۳۸۴	۳،۶۵

بر اساس نتایج عددی حاصل از خروجی کد، مقایسه دو روش یادگیری Q و یادگیری Q دوگانه در دو محیط ۴-حرکتی و ۶-حرکتی نشان دهنده عملکرد متفاوت این الگوریتم‌ها است. در محیط ۴-حرکتی، یادگیری Q با میانگین پاداش ۹۴,۵۷ و دقت موقعیتی ٪۵۵,۳۴، عملکرد به‌طور معناداری (p=0.000000) بهتر از یادگیری Q دوگانه با میانگین پاداش ۸۱,۹۰ و دقت ٪۴۷,۳۰ بود. هرچند یادگیری Q دوگانه در این محیط با ۲۸,۳۵ برخورد در هر اپیزود، برخورد کمتری نسبت به روش معمولی (۳۵,۸۶ برخورد) داشت. در مقابل، در محیط پیچیده‌تر ۶-حرکتی، تفاوت آماری معناداری بین دو روش مشاهده نشد (p=0.079261) و هر دو با دقت موقعیتی حدود ٪۷۹ و نرخ پیشرفت ٪۰,۶ عملکرد بسیار مشابهی ارائه کردند، با این تفاوت که و یادگیری Q دوگانه با میانگین پاداش ۱۴۷,۳۷ و دقت ٪۷۹,۶۱ اندکی برتر بود.

در جمع‌بندی نهایی، همان‌طور که هدف مطالعه نشان می‌دهد و یادگیری Q دوگانه در محیط ۶-حرکتی به عنوان روش بهینه انتخاب شد (با امتیاز ۰,۷۶۲). این روش نه تنها با دقت موقعیتی ۷۹,۶۱٪ بهترین عملکرد را داشت، بلکه با ۳۲,۵۶ برخورد در هر اپیزود، تعادل بهتری بین دقت و ایمنی ایجاد کرد. همچنین تحلیل همگرایی نشان داد و یادگیری Q دوگانه در محیط ۶-حرکتی با ۸,۲ اپیزود سریع‌تر همگرا شده است. بنابراین می‌توان نتیجه گرفت که در محیط‌های پیچیده‌تر با فضای عملیاتی بزرگ‌تر (۶-حرکتی)، یادگیری Q دوگانه به عنوان روش پایدارتر و کارآمدتر توصیه می‌شود، درحالی‌که در محیط‌های ساده‌تر (۴-حرکتی) یادگیری Q معمولی می‌تواند گزینه مناسب‌تری باشد.

نتایج نشان می‌دهد که در روش بهینه یادگیری Q دوگانه، میانگین برخورد در هر اپیزود ۳۲,۵۵ است. با توجه به اینکه هر اپیزود حداکثر ۱۰۰ گام و ۱۰ ربات دارد، این میزان برخورد قابل توجه است. بررسی داده‌ها نشان می‌دهد در اپیزودهایی که موفقیت کامل حاصل نشده، میانگین برخوردها به‌طور محسوسی بالاتر است (اغلب بیش از ۵۰ برخورد). با توجه به وجود ۲ مانع متحرک با رفتار تصادفی (احتمال ۵٪ تغییر جهت ناگهانی)، این موانع نقش اصلی را در شکست‌ها ایفا می‌کنند. برخلاف موانع ثابت که الگوی مشخصی دارند، موانع متحرک می‌توانند به‌طور پیش‌بینی‌نشده وارد مسیر ربات شده و با اعمال جریمه سنگین (۱۰-)، ربات را از مسیر بهینه منحرف کنند. تخمین زده می‌شود حدود ۶۰-۷۰٪ از شکست‌ها مستقیماً ناشی از برخورد با موانع متحرک باشد.

با بررسی روند یادگیری در خروجی کد (مقایسه اپیزودهای اولیه و نهایی)، مشاهده می شود که در ۲۰۰ اپیزود اول، نرخ موفقیت به طور میانگین حدود ۵۰-۶۰٪ بوده، در حالی که در ۲۰۰ اپیزود پایانی به حدود ۸۰-۹۰٪ رسیده است. همچنین تحلیل همگرایی نشان می دهد که روش یادگیری Q دوگانه با ۶ حرکت به‌طور متوسط پس از ۸,۲ اپیزود به همگرایی می‌رسد (یعنی به نرخ موفقیت بالای ۵۰٪ دست می‌یابد). با توجه به اینکه ۸۰۰ اپیزود اجرا شده و منحنی یادگیری هنوز در حال بهبود است، افزایش اپیزودها به ۱۲۰۰-۱۰۰۰ می تواند نرخ موفقیت را به حدود ۸۵-۸۷٪ برساند. با این حال، دستیابی به ۱۰۰٪ موفقیت به دلیل ماهیت غیرقطعی موانع متحرک و محدودیت حداکثر ۱۰۰ گام، عملاً غیرممکن است و همواره شانس کمی (حدود ۵-۱۰٪) برای برخورد تصادفی با مانع متحرک وجود خواهد داشت.

نرخ موفقیت حدود ۷۹٪ در شرایط حضور ۲ مانع متحرک تصادفی، ۱۲ مانع ثابت، ۱۰ ربات همزمان و حداکثر ۱۰۰ گام، از نظر عملیاتی عملکرد بسیار خوب و قابل قبولی محسوب می شود. علت اصلی ۲۱٪ شکست، برخوردهای پیش‌بینی‌نشده با موانع متحرک است. افزایش اپیزودها تا ۱۲۰۰ می تواند نرخ موفقیت را تا حدود ۸۷٪ بهبود بخشد، اما حذف کامل شکست‌ها در این محیط غیرقطعی، شدنی نیست.

در خصوص تحلیل رفتار الگوریتم‌ها در شرایط افزایش مقیاس محیط و تعداد موانع، با افزایش ابعاد شبکه از ۲۰×۲۰ به ۱۰۰×۱۰۰ و افزایش متناظر موانع ثابت از ۱۲ به ۴۰ و موانع متحرک از ۲ به ۱۰، کاهش چشمگیری در عملکرد همه الگوریتم‌ها مشاهده گردید. موفقیت بهترین روش

الگوریتم یادگیری Q دوگانه (با ۶ حرکت) از حدود ۷۹٪ به حدود ۱۹٪ کاهش یافت. علت اصلی این کاهش، افزایش نمایی فضای حالت (از ۴۰۰ به ۱۰،۰۰۰ حالت) و عدم بازدید کافی از حالت‌ها در ۸۰۰ اپیزود است. همچنین افزایش موانع متحرک، عدم قطعیت محیط را افزایش داده و افزایش موانع ثابت، مسیرهای قابل عبور را محدود نموده است. بر خلاف محیط کوچک که تفاوت معناداری بین روش‌های ۴ و ۶ حرکتی مشاهده نشد، در محیط بزرگ، روش ۶-حرکتی برتری معنی‌داری ($p<0.05$) نسبت به روش ۴-حرکتی نشان داد. همچنین هیچیک از روش‌ها در ۸۰۰ اپیزود به همگرایی نرسیدند که نشان‌دهنده نیاز به اپیزودهای بیشتر یا استفاده از روش‌های تقریب تابع است.

در پژوهش حاضر، صرفاً از الگوریتم‌های جدول‌محور یادگیری Q و یادگیری Q دوگانه استفاده شده است. دلیل اصلی این انتخاب، تمرکز مطالعه بر تحلیل مبانی نظری رفتار این الگوریتم‌ها در شرایط عدم قطعیت محیطی و در یک فضای حالت گسسته با ابعاد کوچک (شبکه ۲۰×۲۰ با ۴۰۰ حالت) می‌باشد. چنین فضای حالتی برای روش‌های جدول‌محور کاملاً قابل مدیریت است و امکان استخراج دقیق ماتریس Q بدون نیاز به تقریب تابع فراهم می‌آورد. به‌کارگیری روش‌های پیشرفته‌تری مانند شبکه‌های Q عمیق (DQN) در این ابعاد، به دلیل وجود خطای تقریب و وابستگی شدید به فراپارامترهای شبکه عصبی، ارزیابی‌های آماری پایدار و قابل مقایسه با روش‌های جدول‌محور را دشوار می‌سازد. افزون بر این، هدف اصلی پژوهش مقایسه مستقیم اثر مکانیزم کاهش بیش‌برآورد در یادگیری Q دوگانه در برابر نسخه استاندارد آن بوده است؛ این مقایسه تنها در صورت حذف خطای تقریب و استفاده از ساختار جدول‌محور، از خلوص نظری برخوردار خواهد بود.

از منظر کیفی، انتظار می‌رود روش‌های یادگیری عمیق تقویتی و الگوریتم‌های چندعاملی بازیگر-نقاد مانند بهینه‌سازی سیاست نزدیک‌شونده چندعاملی[۲۵] مزایای قابل توجهی را در فضاهای حالت و عمل با ابعاد بسیار بزرگ یا محیط‌های دارای مشاهدات پیوسته ارائه دهند. الگوریتم‌های چندعاملی مانندMAPPO نیز با بهره‌گیری از یادگیری متمرکز و اجرای غیرمتمرکز می‌توانند هماهنگی مؤثرتری میان عوامل ایجاد کنند. با وجود این، پیاده‌سازی این روش‌ها نیازمند حجم داده، توان محاسباتی و زمان اجرای بسیار بیشتری است. از آنجا که پژوهش حاضر بر شفافیت آماری، تکرارپذیری مبتنی بر شبیه‌سازی مونت‌کارلو و فضای حالت کوچک تأکید دارد، استفاده از روش‌های عمیق و چندعاملی در این مطالعه لحاظ نشده است. با این حال، مقایسه سیستماتیک روش‌های جدول‌محور با رویکردهای یادگیری عمیق و چندعاملی در محیط‌های با مقیاس بزرگ، یکی از مسیرهای ضروری برای پژوهش‌های آتی محسوب می‌شود (همان‌گونه که در پیشنهادهای شماره ۴ و ۵ مقاله نیز اشاره شده است.).

تحلیل حساسیت فرا پارامترهای یادگیری

به منظور بررسی تأثیر تغییر فراپارامترهای کلیدی بر عملکرد یادگیری، تحلیل حساسیت بر روی سه پارامتر اصلی نرخ یادگیری (α) ، ضریب تنزیل (γ) و نرخ اکتشاف (ε) انجام پذیرفت. مطابق با رویکردهای استاندارد در تحلیل حساسیت یادگیری تقویتی، تغییرات هر پارامتر به صورت مجزا و با فرض ثابت بودن سایر پارامترها

در مقادیر پایه (α=0.1 ، γ=0.95و ε=0.1) اعمال گردید. هر پیکربندی شامل ۲۰۰ اجرای مستقل مونت‌کارلو به منظور اطمینان از پایایی آماری نتایج بود. برای کمی‌سازی حساسیت عملکرد نسبت به تغییرات هر پارامتر، شاخص‌های زیر تعریف شد:

- تغییرات نسبی[۲۶] برای هر معیار عملکرد (مانند نرخ موفقیت یا میانگین پاداش) به صورت زیر محاسبه گردید:

$$RD_P^{(x_i)} = \frac{|f(x_i)-f(x_{base})|}{f(x_{base})} \qquad (۱۸)$$

که در آن $f(x_i)$ مقدار معیار عملکرد در پیکربندی جدید و $f(x_{base})$ مقدار آن در پیکربندی پایه می‌باشد.

-شاخص حساسیت[۲۷] با استفاده از رابطه زیر تعریف شد:

$$SI = \frac{\Delta J/J_{nom}}{\Delta p/p_{nom}} = \frac{(J-J_{nom})/J_{nom}}{(p-p_{nom})/p_{nom}} \qquad (۱۹)$$

که در آن J مقدار معیار عملکرد، J_{nom} مقدار عملکرد در نقطه پایه، p مقدار جدید فراپارامتر و p_{nom} مقدار پایه فراپارامتر است. مقادیر SI نشان‌دهنده حساسیت زیاد ($SI > 1$)، متوسط ($0.5 < SI < 1$) و کم ($SI \leq 0.5$) می‌باشند.-

نرمال‌سازی کارایی[۲۸] به منظور مقایسه یکپارچه عملکرد در معیارهای مختلف، هر معیار با استفاده از رابطه زیر در بازه [۰،۱] نرمال گردید:

$$PN_i = \frac{J_i-J_{min}}{J_{max}-J_{min}} \qquad (۲۰)$$

و سپس امتیاز کلی به عنوان میانگین هندسی معیارهای کلیدی شامل دقت موقعیتی، میانگین پاداش و میانگین برخورد محاسبه شد:

$$OS = \sqrt[3]{PN_{success} \cdot PN_{reward} \cdot PN_{collision}} \qquad (۲۱)$$

نتایج تحلیل حساسیت در جدول (۵) ارائه شده است. مشاهدات اصلی حاکی از آن است که نرخ یادگیری (α) بیشترین تأثیر را بر سرعت همگرایی و پایداری یادگیری دارد، به‌طوری که کاهش α به ۰،۰۱ سرعت همگرایی را به‌طور چشمگیری کاهش داده (همگرایی پس از ۳۵۰ اپیزود در مقایسه با ۱۲۰ اپیزود در حالت پایه)، در حالی که افزایش α به ۰،۳ منجر به نوسانات شدید و کاهش نرخ موفقیت نهایی به حدود ۶۸٪ می‌گردد. در مقابل، ضریب تنزیل (γ) تأثیر مستقیمی بر کیفیت سیاست نهایی دارد؛ مقدار پایین γ=0.9 رفتار کوتاه‌بینانه ایجاد کرده و نرخ موفقیت نهایی را به حدود ۷۱٪ کاهش می‌دهد، در حالی که γ=0.99 سیاست بلندمدت بهتری شکل داده و نرخ موفقیت را تا ۸۳٪ افزایش می‌دهد .(SI=1.96) نرخ اکتشاف اولیه (ε) تعادل بین اکتشاف و بهره‌برداری را کنترل می‌کند؛ مقدار ε=0.3 منجر به اکتشاف بیش از حد و افزایش برخوردها (میانگین ۴۱،۲ برخورد در هر اپیزود) می‌شود، در حالی که ε=0.05 به دلیل کاهش ناکافی اکتشاف، الگوریتم را در بهینه‌های محلی گرفتار کرده و نرخ موفقیت را به حدود ۷۲٪ کاهش می‌دهد. یافته فوق با نتایج مطالعات پیشین مبنی بر وجود رابطه جبرانی بین α و γ در یادگیری Q همسو می‌باشد، به‌طوری که مقادیر نامناسب α را می‌توان تا حدودی با تنظیم γ جبران کرد. در مجموع، پیکربندی پایه(α=0.1 ، γ=0.95، ε=0.1)از منظر توأمان دقت، سرعت و پایداری یادگیری، مناسب‌ترین تعادل را ارائه می‌دهد.

α	α	γ
۰،۲	۰،۳	۰،۹۰
۷۵،۵	۶۸،۹	۷۱،۳
۱۴۳،۱	۱۳۱،۲	۱۲۸،۹
۳۶،۸	۴۳،۵	۲۹،۸
۹۰	۷۵	۱۴۰
۱۳۳،۳	۱۶۰،۰	۸۵،۷
۰،۰۵۲	۰،۱۳۴	۰،۱۰۴
۰،۳۹	۰،۴۰	۰،۶۹
همگرایی سریع اما نوسان‌های انتهایی و برخورد بی‌شتر	نوسانات شدید، عدم پایداری و کاهش قابل‌توجه موفقیت	کوتاه‌بینانه عمل کرده، کیفیت سیاست نهایی پایین‌تر است

جدول ۵. تحلیل حساسیت فرایارامترهای یادگیری (α ،γ ،ε)

پارامتر	پایه	α	α	α
مقدار	$\alpha=0.1,\gamma=0.95,\varepsilon=0.1$	۰،۰۱	۰،۰۵	۰،۱
نرخ موفقیت نهایی(%)	۷۹،۶	۶۸،۳	۷۴،۲	۷۹،۶
میانگین پاداش نهایی	۱۴۷،۴	۱۲۲،۵	۱۳۵،۸	۱۴۷،۴
میانگین برخورد در هر اپیزود	۳۲،۶	۲۸،۱	۳۰،۹	۳۲،۶
تعداد اپیزود تا همگرایی	۱۲۰	۳۵۰	۱۸۰	۱۲۰
سرعت همگرایی نسبی(%)	۱۰۰،۰	۳۴،۲	۶۶،۷	۱۰۰،۰
آر.دی موفقیت	۰،۰۰۰	۰،۱۴۲	۰،۰۶۸	۰،۰۰۰
اس.آی موفقیت	—	۰،۵۶	۱،۱۷	—
تفسیر عملکردی	تعادل بهینه بین سرعت، دقت و پایداری	همگرایی بسیار کند؛ اکتشاف ناکافی در ابتدا	سرعت همگرایی بهبود یافته اما حساس به نوسان	مقدار بهینه برای این مسئله

γ	ε₀	
۰٫۹۵	۰٫۰۵	کی‌فی‌ت سی‌است بالاتر اما با برخورد بی‌شتر و همگرای‌ی کندتر
۷۹٫۶	۷۲٫۴	اکتشاف ناکافی، گرفتار شدن در بهی‌نه محلی
۱۴۷٫۴	۱۳۳٫۵	
۳۲٫۶	۲۸٫۳	
۱۲۰	۹۵	
۱۰۰٫۰	۱۲۶٫۳	
۰٫۰۰۰	۰٫۰۹۰	
—	۰٫۱۶	مقدار بهی‌نه برای ای‌ن مسئله

γ	ε₀	
۰٫۹۵	۰٫۱	مقدار بهی‌نه برای ای‌ن مسئله
۷۹٫۶	۷۲٫۴	
۱۴۷٫۴	۱۳۳٫۵	
۳۲٫۶	۲۸٫۳	
۱۲۰	۹۵	
۱۰۰٫۰	۱۲۶٫۳	
۰٫۰۰۰	۰٫۰۹۰	
—	۰٫۱۶	کی‌فی‌ت سی‌است بالاتر اما با برخورد بی‌شتر و همگرای‌ی کندتر

γ	ε₀	
۰٫۹۹	۰٫۳	اکتشاف بی‌ش از حد، برخوردهای زی‌اد، همگرای‌ی دی‌رهنگام
۸۳٫۲	۷۳٫۴	
۱۵۶٫۸	۱۳۵٫۲	
۳۶٫۴	۴۱٫۲	
۱۵۵	۱۶۰	
۷۷٫۴	۷۵٫۰	
۰٫۰۴۵	۰٫۰۷۸	
۱٫۹۶	۰٫۳۰	اکتشاف بی‌ش از حد، افت اندک در عملکرد

γ	ε₀	
۰٫۹۵	۰٫۲	اکتشاف نسبتاً بی‌ش از حد، افت اندک در عملکرد
۷۹٫۶	۷۶٫۸	
۱۴۷٫۴	۱۴۱٫۹	
۳۲٫۶	۳۶٫۹	
۱۲۰	۱۳۵	
۱۰۰٫۰	۸۸٫۹	
۰٫۰۰۰	۰٫۰۳۵	
—	۰٫۴۵	مقدار بهی‌نه برای ای‌ن مسئله

نتایج تحلیل حساسیت نشان می‌دهد که نرخ یادگیری (α) بیشترین تأثیر را بر سرعت همگرایی دارد، به‌طوری‌که کاهش آن به ۰٫۰۱ منجر به کاهش ۱۱٫۳٪ در نرخ موفقیت و افزایش ۱۹۲٪ در زمان همگرایی می‌شود (SI=0.56) ، در حالی‌که افزایش آن به ۰٫۳ نوسانات شدید و افزایش ۳۳٫۴٪ در میانگین برخوردها را به همراه دارد. ضریب تنزیل (γ) تأثیر مستقیمی بر کیفیت سیاست نهایی دارد؛ مقدار $\gamma=0.99$ ضمن افزایش ۳٫۶٪ نرخ موفقیت نسبت به حالت پایه، به دلیل دید بلندمدت‌تر، شاخص حساسیت بالایی (SI=1.96)از خود نشان می‌دهد، هرچند با افزایش ۱۱٫۷٪ برخوردها همراه است. نرخ اکتشاف اولیه (ε)نیز تعادل بین اکتشاف و بهره‌برداری را کنترل می‌کند؛ مقادیر بسیار پایین (۰٫۰۵) الگوریتم را

در بهینه‌های محلی گرفتار کرده و مقادیر بالا (۰,۳) اکتشاف بیش از حد و افزایش ۲۶,۴٪ برخوردها را سبب می‌شوند. در مجموع، پیکربندی پایه (α=0.1 ، γ=0.95، ε=0.1) از نظر توأمان دقت موقعیتی (۷۹,۶٪)، میانگین پاداش (۱۴۷,۴) و سرعت همگرایی (~۱۲۰ اپیزود) بهترین تعادل را ارائه داده و نسبت به سایر پیکربندی‌ها از حساسیت کمتری برخوردار است؛ بنابراین می‌توان آن را به عنوان انتخاب بهینه برای مسئله مسیریابی گروهی در محیط پویا معرفی کرد.

۷. نتیجه‌گیری

هدف اصلی این پژوهش، انجام یک ارزیابی عمیق و مبتنی بر شواهد آماری از عملکرد الگوریتم‌های یادگیری Q و یادگیری دوگانه Q در حل مسئله هدایت گروهی ربات‌ها در محیطی پویا و غیرقطعی است. برای نیل به این هدف کلی، اهداف ویژه‌ای شامل طراحی یک محیط شبیه‌سازی گسسته مجهز به موانع متحرک با الگوی رفتاری تصادفی، به کارگیری روش شبیه‌سازی مونت‌کارلو برای اطمینان از پایایی و تکرارپذیری نتایج، تعریف و اندازه‌گیری مجموعه‌ای از معیارهای کمی کلیدی مانند نرخ موفقیت، میانگین برخورد، کارایی مسیر و سرعت همگرایی و در نهایت اجرای تحلیل‌های آماری پیشرفته برای تعیین معناداری و عمق تفاوت‌های مشاهده شده بین دو الگوریتم دنبال می‌شوند. اگرچه الگوریتم‌های یادگیری Q و یادگیری Q دوگانه به‌طور گسترده در حوزه‌های مختلف یادگیری تقویتی مطالعه شده‌اند، اما مقایسه سیستماتیک، کمی و مبتنی بر تحلیل‌های آماری پیشرفته این دو روش در هدایت بلادرنگ ناوگان ربات‌ها در محیط‌های پویا با موانع متحرک تصادفی، شکافی قابل توجه در این حوزه پژوهشی محسوب می‌شود. به ویژه، ارزیابی همزمان این الگوریتم‌ها با به‌کارگیری شبیه‌سازی مونت‌کارلو برای اطمینان از پایایی آماری و تحلیل چند معیاره شاخص‌هایی چون نرخ موفقیت، میانگین برخورد، پایداری همگرایی، کارایی مسیر و مقاومت در برابر عدم قطعیت محیطی، تاکنون در قالب یک مطالعه یکپارچه کمتر مورد توجه قرار گرفته است. پر کردن این خلأ برای انتخاب الگوریتم بهینه بر اساس اولویت‌های عملیاتی متفاوت (مانند ایمنی در مقابل سرعت یا پایداری در مقابل دقت) ضروری است. هدایت هماهنگ و ایمن ربات‌ها در محیط‌های عملیاتی پویا و غیرقطعی، به ویژه در حضور موانع متحرک با رفتار تصادفی به یک چالش پیچیده در رباتیک و سیستم‌های خودران تبدیل شده است. موفقیت در چنین مأموریت‌هایی مستلزم الگوریتم‌هایی است که توانایی یادگیری سیاست‌های بهینه، سازگاری بلادرنگ با تغییرات محیطی و تصمیم‌گیری جمعی را داشته باشند. در میان رویکردهای مختلف، الگوریتم‌های یادگیری تقویتی به عنوان چارچوبی قدرتمند برای حل مسائل تصمیم‌گیری ترتیبی مطرح شده‌اند. با این حال، انتخاب بین گونه‌های مختلف این الگوریتم‌ها به‌طور خاص بین یادگیری Q و یادگیری Q دوگانه برای کاربرد در هدایت ناوگان ربات‌ها در محیط های پویا، نیازمند بررسی تجربی دقیق و تحلیل‌های آماری نظام‌مند است. این مقاله بیان کننده مقایسه‌ای جامع است که عملکرد این دو الگوریتم پایه را هم‌زمان از منظر معیارهای چندگانه عملکردی، پایداری یادگیری و قابلیت اطمینان آماری در شرایط شبیه‌سازی‌شده

نزدیک به واقعیت با تأکید بر حضور موانع متحرک تصادفی مورد ارزیابی قرار می دهد. فرضیه اصلی این تحقیق بر این است که یادگیری Q دوگانه به دلیل معماری دوگانه و مکانیزم کاهش بیش برآورد، در محیط‌های پویا و غیر قطعی عملکرد پایدارتر و ایمن‌تری نسبت به یادگیری Q استاندارد ارائه می‌دهد که این برتری به صورت کاهش آماری معنادار در میانگین برخوردها و واریانس کمتر در خروجی‌های اجراهای مونت‌کارلو منجر شد. در مقابل، یادگیری Q در برخی پارامترهای نقطه‌ای مانند سرعت اولیه همگرایی یا پاداش میانگین در سناریوهای خاص عملکرد بهتری نشان دهد. همچنین انتظار می‌رود که افزایش درجه آزادی عمل (۴از عمل به ۶ عمل)، عملکرد هر دو الگوریتم را بهبود بخشد، اما سود حاصل از این افزایش برای یادگیری Q دوگانه بارزتر می باشد.

در ادامه چند مسیر پژوهشی برای ادامه کار پیشنهاد شده است که عبارتند از:

۱. توسعه روش‌های مسیریابی آنلاین در محیط‌های کاملاً پویا با استفاده از یادگیری تقویتی
۲. مقایسه سیستماتیک الگوریتم‌های یادگیری تقویتی با روش‌های کلاسیک مسیریابی مانند آاستار
۳. طراحی الگوریتم‌های مسیریابی گروهی آنلاین برای ناوگان ربات‌های همکار در محیط‌های ناشناخته
۴. بهره‌گیری از شبکه‌های عصبی عمیق در محیط‌هایی با مقیاس بزرگ
۵ .ترکیب شبکه‌های عصبی گرافی با یادگیری تقویتی چندعاملی برای بهبود هماهنگی در مسیریابی گروهی

۸. منابع

[1] P. Guo, R. Zhang, and B. Xu, "Safety Separation Distance Design for UAV Formation Based on System Performance," IEEE Transactions on Aerospace and Electronic Systems, pp. 1–16, 2025.

[2] E. C. Ozkat, "Vibration data-driven anomaly detection in UAVs: A deep learning approach," Engineering Science and Technology, an International Journal, vol. 54, p. 101702, 2024/06/01/ 2024.

[3] M. E. Elshaar, M. R. Elbalshy, A. Hussien, and M. Abido, "Path Planning in a dynamic environment using Spherical Particle Swarm Optimization," in 2024 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8, July 2024.

[4] X. Olaz, D. Alaez, M. Prieto, J. Villadangos, and J. J. Astrain, "Quadcopter neural controller for take-off and landing in windy environments," Expert Systems with Applications, vol. 225, p. 120146, 2023.

[5] X. Ai, Y. Zhang, and Y.-Y. Chen, "Spherical Formation Tracking Control of Non-Holonomic UAVs with State Constraints and Time Delays," Aerospace, vol. 10, no. 2, 2023.

[6] S. Poudel, M. Y. Arafat, and S. Moh, "Bio-Inspired Optimization-Based Path Planning Algorithms in Unmanned Aerial Vehicles: A Survey," Sensors, vol. 23, no. 6, 2023.

[7] X. Wang, L. Wang, and Y. Hu, "Self-confirming Q-learning on unknown networks," Chaos, Solitons & Fractals, vol. 203, p. 117598, 2026.

[8] M. I. A. Shah, M. E. Cruz Victorio, M. Duffy, E. Barrett, and K. Mason, "Peer-to-peer energy trading in dairy farms using multi-agent reinforcement learning," Applied Energy, vol. 402, p. 127041, 2026.

[9] S. Lin, J. Wang, B. Huang, X. Kong, and H. Yang, "Bio particle swarm optimization and reinforcement learning algorithm for path planning of automated guided vehicles in dynamic industrial environments," Scientific Reports, vol. 15, no. 1, p. 463, 2025.

[10] C. Zhang, K. K. H. Ng, Z. Jin, S. Yao, and Y. Qin, "Q-learning-driven exact and meta-heuristic algorithms for the robust gate assignment problem," Advanced Engineering Informatics, vol. 67, 2025.

[11] B. Q. A. Nguyen, N. T. Dang, T. T. Le, and P. N. Dao, "On-policy and Off-policy Q-learning algorithms with policy iteration for two-wheeled inverted pendulum systems," Robotics and Autonomous Systems, p. 105111, 2025.

[12] C. Zhang, K. Ng, Z. Jin, S. Yao, and Y. Qin, "Q-learning-driven exact and meta-

heuristic algorithms for the robust gate assignment problem," Advanced Engineering Informatics, vol. 67, p. 103551, 2025.
[13] C. Sun, L. Hou, S. Yu, and J. Shu, "HQA: Hybrid Q-Learning and AODV Multi-Path Routing Algorithm for Flying Ad-Hoc Networks," Vehicular Communications, p. 100947, 2025.
[14] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in Proceedings of the AAAI conference on artificial intelligence, vol. 30, no. 1, 2016.
[15] X. Guo, Q.-K. Pan, W. Zhang, Z.-H. Miao, X.-L. Jing, and H.-Y. Sang, "A Q-learning-assisted memetic algorithm for joint vehicle scheduling problem for harvesting and transportation in smart agriculture," Swarm and Evolutionary Computation, vol. 96, p. 102007, 2025.
[16] Z. Duan, Y. Zhang, Z. Shao, Z. Xu, and Z. Xiang, "Advanced robot path planning on rough terrain: A Q-learning-based multi-objective PSO algorithm," Applied Soft Computing, p. 113798, 2025.
[17] Z. Li, Y. Wang, Y. Han, K. Gao, and J. Li, "Q-Learning-Driven Accelerated Iterated Greedy Algorithm for Multi-Scenario Group Scheduling in Distributed Blocking Flowshops," Knowledge-Based Systems, vol. 317, p. 113424, 2025.
[18] A. Giuffrida, N. Basilico, and F. Amigoni, "An empirical evaluation of learning-based multi-agent path finding algorithms in warehouse environments," Robotics and Autonomous Systems, p. 105149, 2025.
[19] J. Gao, Y. Li, X. Shi, and K. Yan, "Customer satisfaction optimization in due-aware multi-agent path finding," Information Sciences, p. 122782, 2025.
[20] S. Venu and M. Gurusamy, "A Comprehensive Review of Path Planning Algorithms for Autonomous Navigation," Results in Engineering, p. 107750, 2025.
[21] S. Aslan and S. Demirci, "An immune plasma algorithm with Q-learning based pandemic management for path planning of unmanned aerial vehicles," Egyptian Informatics Journal, vol. 26, p. 100468, 2024.
[22] Y. Wang and G. Wang, "An adaptive and efficient path planning algorithm for UAV navigation in complex environments," Computers & Operations Research, vol. 185, p. 107296, 2026.
[23] X. Ni, W. Hu, Q. Fan, Y. Cui, and C. Qi, "A Q-learning based multi-strategy integrated artificial bee colony algorithm with application in unmanned vehicle path planning," Expert Systems with Applications, vol. 236, p. 121303, 2024.
[24] E. Mohammad, M. Jahromi, J. Pirkandi, M. Khazaee, and M. Mahmoodi, "Development of an intelligent jet engine controller using a model-based deep deterministic policy gradient technique,", Journal of Mechanical Engineering and Sciences, vol. 19, no. 3, pp. 10739 – 10755, 2025.

۹. پی‌نوشت‌ها

[1] Reinforcement Learning
[2] Q-Learning
[3] Double Q-Learning
[4] Action
[5] Deep Q-Learning (DQN)
[6] Memetic Algorithms
[7] Nonholonomic constraints
[8] State
[9] Reinforcement learning (RL)
[10] Rewards
[11] Bellman Equation
[12] Action-State
[13] Temporal Difference (TD) Update
[14] Temporal Difference (TD)
[15] Episode
[16] TD-Update
[17] Exploration
[18] Exploitation
[19] Double Q-Learning (DQ)
[20] Overestimation
[21] ε-Greedy Policy
[22] d-Cohen
[23] Levin
[24] Action
[25] Multi-Agent Proximal Policy Optimization (MAPPO)
[26] Relative Deviation (RD)
[27] Sensitivity Index (SI)
[28] Performance Normalization (PN)