فصلنامه مهندسی حمل و نقل، دوره ۱۷، شماره ۲، زمستان ۱۴۰۴، ص. ۵۳۵۸-۵۳۳۳
مقاله علمی-پژوهشی

کنترل شتاب طولی و جانبی خودروی بدون سرنشین با استفاده از یادگیری تقویتی عمیق

محسن ابراهیمی، دانشجوی دکتری، گروه مهندسی برق، دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران
فیروز اللهوردی زاده (مسئول مکاتبات)، استادیار، گروه مهندسی برق، دانشکده مهندسی برق، دانشگاه صنعتی مالک اشتر، تهران، ایران
E-mail: f_alahverdizadeh@mut.ac.ir
عبدالرضا کاشانی نیا، استادیار، گروه مهندسی برق، دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر، تهران، ایران

دریافت: ۱۴۰۳/۱۱/۲۱ پذیرش: ۱۴۰۴/۰۶/۱۶

چکیده

کنترل دقیق شتاب طولی و جانبی در خودروهای بدون سرنشین یکی از چالش‌های کلیدی در توسعه سیستم‌های خودران ایمن و کارآمد به شمار می‌رود. با توجه به رشد سریع فناوری‌های خودران و نیاز به بهبود عملکرد این سیستم‌ها در شرایط مختلف جاده‌ای، پژوهش در زمینه کنترل هوشمند این خودروها اهمیت ویژه‌ای یافته است. در این مقاله، یک روش نوآورانه برای کنترل شتاب طولی و جانبی خودروهای بدون سرنشین مبتنی بر یادگیری تقویتی ارائه شده است. روش پیشنهادی با بهره‌گیری از معماری تفکیک‌شده، پیچیدگی محاسباتی را کاهش داده و امکان استفاده از الگوریتم‌های یادگیری تقویتی پیشرفته‌تر را فراهم می‌کند. در مرحله اولیه، یک عامل یادگیری تقویتی واحد برای کنترل هم‌زمان شتاب طولی و جانبی طراحی و آموزش داده می‌شود. سپس، به‌منظور بهبود کارایی، عامل‌های یادگیری تقویتی به دو بخش مستقل برای کنترل طولی و جانبی تفکیک‌شده و به‌صورت جداگانه آموزش داده می‌شوند. نتایج شبیه‌سازی‌ها نشان می‌دهد که این تفکیک نه‌تنها سرعت همگرایی فرآیند آموزش را افزایش می‌دهد، بلکه دقت عملکرد سیستم کنترل را نیز به‌طور قابل توجهی بهبود می‌بخشد. جداسازی عامل‌ها، کاهش حدود ۲ ساعت از زمان آموزش شبکه‌های عمیق، تقلیل ۳۸,۲ درصدی میانگین خطای سمت و ۱۰,۱ درصدی میانگین خطای فاصله را نتیجه می‌دهد. این یافته‌ها می‌تواند به توسعه سیستم‌های کنترلی پیشرفته‌تر و ایمن‌تر برای خودروهای خودران کمک کند و نقش مهمی در ارتقاء ایمنی و کارایی این فناوری‌ها ایفا نماید.

واژه‌های کلیدی: خودروی بدون سرنشین، شتاب جانبی خودرو، شتاب طولی خودرو، یادگیری تقویتی

۱. مقدمه

جراحات جاده‌ای یکی از عوامل اصلی مرگ‌ومیر در جهان است. گزارش‌های سازمان بهداشت جهانی [۱] نشان می‌دهد که تقریباً ۱,۳ میلیون نفر سالانه در اثر تصادفات جاده‌ای جان خود را از دست می‌دهند [Peličić, Ristić and Radević, 2024]. کاربران آسیب‌پذیر جاده [۲] که شامل عابران پیاده، دوچرخه و موتورسیکلت میشوند بیش از نیمی از این تعداد را به خود اختصاص داده است. مطالعات دیگر نشان می‌دهد که عابران پیاده اکثریت قربانیان VRU را تشکیل می‌دهد [Hamdani, Benamar and Younis, 2020]. مطالعه اخیر گروه متمرکز بر هوش مصنوعی، در رابطه با خودروی خودران [۳]، نشان می‌دهد که آسیب‌های جاده‌ای در حال حاضر علت اصلی مرگ‌ومیر کودکان است و بسیار بیشتر از مرگ‌های ناشی از اچ‌آی‌وی و سل است.

پیشرفت خودروهای خودران [۴] می‌تواند به کاهش مرگ‌ومیر ناشی از تصادفات رانندگی کمک کند و مزایای دیگری نیز به همراه دارد. این خودروها به رانندگان این امکان را می‌دهند که به‌جای تمرکز بر رانندگی، به کارهای دیگر بپردازند. تاریخچه توسعه این فناوری به دهه ۱۹۸۰ برمی‌گردد، زمانی که محققانی مانند ارنست دیکنز از دانشگاه مونیخ و گروه ناولب دانشگاه کارنگی ملون اولین نمونه‌های خودروهای خودران را با قابلیت‌های محدود ایجاد کردند. در اوایل دهه ۱۹۹۰، پروژه‌های مختلفی مانند پروژه اورکا و آزمایش‌های دارپا آغاز شد که به پیشرفت‌های قابل توجهی در این زمینه منجر شد. چالش‌های دارپا در سال‌های ۲۰۰۴ و ۲۰۰۵ نشان داد که خودروهای خودران می‌توانند بدون دخالت انسانی مسافت‌های طولانی را طی کنند. این چالش‌ها موجب تغییر نگرش عمومی نسبت به امکان‌پذیری خودران بودن وسایل نقلیه شد و زمینه را برای ورود شرکت‌هایی مانند گوگل به این حوزه فراهم کرد. تا سال ۲۰۱۰، خودروهای خودران گوگل بیش از ۱۴۰ هزار مایل را در کالیفرنیا طی کردند و نتایج مثبتی را در کاهش مرگ‌ومیر ناشی از تصادفات نشان دادند. پیشرفت‌ها در حسگرها، هوش مصنوعی و ارتباطات بین وسایل نقلیه [۵] به افزایش ایمنی و کارایی این فناوری کمک کرده است [Yusuf, Khan, and Souissi, 2024]. انتظار می‌رود که تا سال ۲۰۲۵، خودروهای خودران به سطح‌های بالاتری از خودمختاری (سطح ۴ و ۵) برسند و همکاری‌های استراتژیک بین شرکت‌ها به تسریع توسعه این فناوری کمک کند. همچنین، استفاده از اینترنت اشیا [۶] در خودروها می‌تواند ارتباطات را بهبود بخشد و تصمیم‌گیری‌های سریع‌تر و هوشمندانه‌تری را ممکن سازد. درنهایت، آینده خودروهای خودران نویدبخش تحولی در نحوه سفر و حمل‌ونقل خواهد بود که می‌تواند ایمنی و راحتی را برای کاربران افزایش دهد.

در [Singh, 2015] گزارش شده است که ۹۰ درصد از تمام تصادفات رانندگی ناشی از خطاهای انسانی تخمین زده می‌شود. پیشرفت‌های اخیر در سیستم‌های حمل‌ونقل هوشمند [۷]، سیستم‌های محاسباتی و هوش مصنوعی باعث افزایش ایمنی شده و راه را برای معرفی گسترده خودروهای خودران هموار و فرصت‌های جدیدی را برای جاده‌های هوشمند، ایمنی ترافیک هوشمند و راحتی مسافران باز کرده است. محققان تخمین می‌زنند که ۸ میلیون خودروهای خودران تا سال ۲۰۲۵ به جاده‌ها برسند [Bay 2021].

یک وسیله نقلیه خودران می‌تواند محیط خود را تشخیص دهد و بدون دخالت انسان تصمیم‌گیری کند [Ebrahimi and Nasrollahi, 2025]. وسایل نقلیه خودران به‌طور مشترک اطلاعات را با یکدیگر، با زیرساخت‌های کنار جاده به اشتراک می‌گذارند [Lamssaggad e tal 2021]. خودروهای خودران با تکیه بر ارتباطات خودرو، پیام‌های ایمنی، شرایط ترافیکی و پیام‌های هشداردهنده در صورت ترافیک یا تصادف را مبادله می‌کنند.

شبکه‌های عصبی یک رویکرد برای تقلید از هوش انسانی و قادر به یادگیری حجم عظیمی از داده‌ها است. آن‌ها طیف وسیعی از کاربردها، از تشخیص تقلب، جداسازی تصویر و ... تا خودروهای خودران را شامل می‌شوند. بااین‌حال، آن‌ها محدودیت‌هایی شامل آسیب‌پذیری در برابر مسئله‌های متخاصم دارند که در آن داده‌ها با هدف ایجاد اشتباه، در مدل آموخته‌شده ارائه می‌شوند. در چند سال گذشته، آن‌ها به یک جزء کلیدی از پروژه‌های بینایی کامپیوتر تبدیل‌شده‌اند [Ebrahimi, 2023]. شبکه عصبی عمیق با استفاده از لایه‌های پردازش متعدد، می‌تواند فرایند یادگیری عمیق را در صنایع مختلف، مانند پزشکی، مهندسی و دیگر زمینه‌ها اعمال نماید.

در ادامه روند مقاله به این صورت است که ابتدا در بخش ۲ به مرور مطالعات پیشین پرداخته می‌شود. در بخش ۳ مدل دینامیکی خودرو در دو کانال طولی و عرضی معرفی و معادلات حاکم بر آن شرح داده می‌شود. سپس درباره مبانی یادگیری تقویتی و نحوه کارکرد آن توضیح داده می‌شود و روابط ریاضی مورد استفاده در آن در بخش ۴ بیان می‌شود. پس از بررسی مفاهیم بحث یادگیری تقویتی در بخش ۵، کاربرد آن در کنترل خودروهای بدون سرنشین گفته می‌شود و روابط ریاضی حاکم بر یادگیری تقویتی در حوزه خودروی بدون سرنشین، سفارشی‌سازی می‌شوند. پس‌ازآن، در بخش ۶ چالشی که این پژوهش به دنبال حل آن است تعریف می‌شود تا در بخش بعدی نحوه مواجه با این چالش توضیح داده شود. روش‌های ارائه شده در بخش ۷ با ارائه دو مثال در محیط متلب شبیه‌سازی شده و نتایج شبیه‌سازی‌ها با یکدیگر مقایسه شده است. در این بخش سعی شده است تا تمامی پارامترها مانیتور شده تا به بررسی دقیق‌تر نتایج شبیه‌سازی کمک کند. سپس در بخش نتیجه‌گیری، نتایج کلی از پژوهش جمع‌بندی شده و بررسی‌های انجام‌شده در بخش‌های قبلی به‌طور جامع شرح داده شده است. همچنین پیشنهادهایی جهت کمک به پژوهش‌های آینده مطرح‌شده است.

۲. ادبیات پژوهش

یادگیری عمیق مجموعه قدرتمندی از روش‌ها برای یادگیری در شبکه‌های عصبی است. درواقع یادگیری عمیق زیرشاخه‌ای از یادگیری ماشین محسوب می‌شود و نورون‌های شبکه‌های عصبی، اسکلت یادگیری عمیق را تشکیل می‌دهند. شبکه‌های عصبی و یادگیری عمیق در حال حاضر بهترین راه‌حل‌ها را برای بسیاری از مشکلات در تشخیص تصویر، تشخیص گفتار و پردازش زبان طبیعی [۸] ارائه می‌دهند. به دلیل توانایی روش‌های یادگیری عمیق و یادگیری تقویتی [۹]، این روش‌ها در بهبود وظایف خودروهای خودران نیز معروف هستند [Mao, Liu and Qu (2024)].

مرجع [Kuutti et al, 2020] روش‌های یادگیری عمیق را برای کنترل خودروهای خودران و عملکرد امیدوارکننده آن در سناریوهای پیچیده مورد بحث قرار داد، نویسندگان نقاط قوت و محدودیت‌های روش‌های یادگیری عمیق موجود را که برای کنترل خودروی خودران اعمال می‌شود، ارائه کردند. بااین‌حال، تمامی جنبه‌های اصلی رانندگی خودران را پوشش ندادند. در [Ma et al, 2020] شیوه‌های فعلی را با استفاده از روش‌های هوش مصنوعی برای خودروهای خودران تجزیه‌وتحلیل شد و چالش‌ها و مسائل مرتبط با اجرای آن‌ها را مورد بحث قرار گرفت. بااین‌حال، آن‌ها بر روی رویکردهای مبتنی بر یادگیری عمیق و یادگیری تقویتی تمرکز نکردند.

در [Claussmann et al, 2019] مروری بر روش‌های برنامه‌ریزی حرکت [۱۰] ارائه شد. تمرکز آن‌ها بر برنامه‌ریزی بزرگراه‌ها بود. آن‌ها در مورد الگوریتم‌های اصلی در برنامه‌ریزی حرکت و کاربردهای آن‌ها در رانندگی بزرگراه بحث کردند. بااین‌حال، به‌طور خاص نقش یادگیری عمیق و یادگیری تقویتی را در این زمینه از خودروهای خودران نشان ندادند. همچنین [Grigorescu et al, 2020] فناوری‌های یادگیری عمیق مورد استفاده در رانندگی خودران را مورد مطالعه قرار داد و نقاط قوت و محدودیت‌های این روش را در درک محیط، برنامه‌ریزی

مسیر، رفتار و کنترل حرکت برجسته کرد که در آن چالش‌های اصلی فعلی در مورد طراحی معماری‌های هوش مصنوعی برای رانندگی خودران را بررسی کردند. این موضوع تا حدود زیادی روش های کنترل قدیمی را تحت تاثیر قرار داده است [Ebrahimi and Asgari, 2021]. در سال ۲۰۲۰، Feng [et al, 2020] سیستم‌های تشخیص و دسته‌بندی اشیا که برای رانندگی خودکار اعمال می‌شود را ارائه کرد. آن‌ها چالش‌ها و سؤالات باز مربوط به این روش‌های تشخیص را ارائه کردند. در سال بعد، [Ning et al, 2021] معماری‌های AI [۱۱] موجود مورد استفاده در رانندگی خودران را ارائه داد و محدودیت‌های این معماری‌ها را به‌طور خلاصه و با مفهوم هوش مصنوعی انسانی (H-AI) [۱۲] را معرفی کرد. همچنین چالش‌های تحقیقاتی باز را ارائه کردند که باید در آینده به آن‌ها پرداخته شود.

این مطالعات نشان می‌دهد که یادگیری عمیق و یادگیری تقویتی به طور گسترده‌ای در حوزه‌های مختلف خودروهای خودران مورد بررسی قرار گرفته‌اند؛ اما به جنبه‌های کلی یا زیرشاخه‌های خاصی مانند تشخیص اشیاء پرداخته‌اند و به صورت عمیق به مسائل کنترلی و مقایسه معماری‌های متفاوت کنترل نپرداخته‌اند. هدف یادگیری تقویتی یافتن یک فرمان کنترلی بهینه (مثلاً تغییر سرعت، ترمز یا شتاب)، با کاوش در محیط به روش تکرار است [Liu and Diao, 2024]. محیط به خودروهای خودران بر اساس رفتار فعلی آن‌ها پاداش می‌دهد تا خطاهای آن‌ها را در آینده تصحیح کند [Li et al, 2023]. یک رویکرد یادگیری تقویتی برای کنترل سرعت خودروهای خودران برای جلوگیری از برخورد سریع با استفاده از گرادیان سیاست قطعی عمیق ارائه کرد. در [Guo, Cheng and Liu, 2020] یک راهبرد کنترل جانبی برای خودروهای خودران مبتنی بر یادگیری تقویتی در یک بزرگراه سه خطه ارائه شد. هدف آن‌ها اجرای ایمن تغییر لاین بود. نتایج نشان داد که جریان ترافیک، در روش ارائه شده آنها بهبود یافته است. سپس [Wu et al, 2020] روش محدودیت سرعت متغیر دیفرانسیل (DVSL) [۱۳] را برای تنظیم سرعت خودروهای خودران در آزادراه با پنج لاین بررسی کرد. DVSL به‌عنوان یک مسئله فرایند مارکوف مدل شده است و شبیه‌ساز SUMO برای آموزش خودروهای خودران برای یادگیری از طریق تعامل با محیط استفاده می‌شود. نتایج آزمایش نشان داد که پیشنهاد آن‌ها ایمنی در بزرگراه را بهبود می‌بخشد. ژانگ به همراه همکاران [Zhang et al,2018] رویکردی را برای کنترل سرعت خودروهای بدون سرنشین بر اساس DQN [۱۴] و یادگیری Q دوگانه مورد استفاده در [Van, Guez and Silver, 2016]، مورد بررسی قرار دادند. باهری به همراه همکاران [Baheri et al, 2020] روشی برای حفظ لاین در رانندگی شهری ارائه کردند. پیشنهاد آن‌ها مشاهدات حالت را از محیط استخراج می‌کند و از یادگیری تقویتی برای آموزش خودروهای خودران در شبیه‌ساز کارلا استفاده می‌کند. خودروهای خودران در دو شهر و شرایط آب‌وهوایی متفاوت با نتایج موفقیت‌آمیز برای وظایف حفظ خطوط شبیه‌سازی‌شده‌اند [Ye et al, 2020]. یک راهبرد تغییر لاین خودکار در بزرگراه‌ها با استفاده از بهینه‌سازی سیاست پروگزیمال (PPO) [۱۵] و یادگیری تقویتی پیشنهاد کرد. با استفاده از حالت‌های وسیله نقلیه و وسایل نقلیه اطراف، خودروهای خودران یاد می‌گیرند که از برخورد اجتناب کنند و مانورهای نرم انجام دهند. نتایج آزمایش‌ها نشان داد که پیشنهاد آن‌ها مانورهای تغییر مسیر را به‌طور مؤثر و ایمن می‌آموزد. در [Toromanoff et al, 2020] یک رویکرد یادگیری تقویتی برای حل شرایط پیچیده (ازجمله حفظ خط مسیر، عابران پیاده و اجتناب از وسایل نقلیه) ارائه می‌شود که از شبیه‌ساز کارلا برای آموزش مدل خود با استفاده از دوربین برای شناسایی رانندگی شهری استفاده شده است. در این روش از DQN برای آموزش خودروهای خودران استفاده شد تا از محیط یاد بگیرند که چگونه موقعیت‌های قبلی را مدیریت کنند. نتایج شبیه‌سازی نشان داد که پیشنهاد آن‌ها برای محیط‌های ناشناخته قابل تعمیم است.

این پژوهش‌ها نشان می‌دهند که یادگیری تقویتی برای کنترل طولی، عرضی و یا ترکیبی، نتایج موفقی داشته است. با این حال، در هیچ یک از این مطالعات به صورت مستقیم به مقایسه معماری های کنترلی مختلف (مانند تک عامله یا چند عامله) و تأثیر آن‌ها بر روند آموزش و عملکرد نهایی پرداخته نشده است. در [Pérez-Gil et al, 2022] به‌طور خاص، الگوریتم‌های یادگیری تقویتی عمیق مانند شبکه DQN و DDPG[۱۶] به‌منظور مقایسه نتایج بین آن‌ها پیاده‌سازی می‌شوند. نتایج به‌دست‌آمده نشان می‌دهد که هم DQN و هم DDPG به هدف می‌رسند، اما DDPG عملکرد بهتری به دست می‌آورد. DDPG مسیرهایی را بسیار شبیه به کنترل‌کننده‌های کلاسیک به‌عنوان LQR انجام می‌دهد. در [Lee et al, 2022] از طریق آموزش تقویتی معکوس (IRL) یک الگوریتم جدید IRL[۱۷] پیشنهاد می‌شود. روش پیشنهادی، به همراه کنترل‌کننده پیش‌بین در شبیه‌ساز کارلا، با حفظ خطوط و تغییر لاین در یک سناریوی چالش‌برانگیز بزرگراه ترافیکی ارزیابی می‌شود.

[Du et al, 2023] نشان می‌دهد که کنترل سیستم تعلیق مبتنی برEK-DDPG، راحتی سواری را در روسازی‌های ناهموار آموزش‌دیده ۲۷,۹۵٪ و ۳,۳۲٪ در مقایسه با کنترل پیش‌بینی مدل و DDPG بهبود می‌بخشد. همچنین، کنترل تعلیق مبتنی بر EK-DDPGراندمان محاسباتی را ۲۲,۹۷٪ در مقایسه با مدل پایه کنترل‌کننده پیش‌بین بهبود می‌بخشد. چن به همراه همکاران [chen et al, 2020]یک یادگیری تقویتی با استفاده از روش جستجوی درخت مونت‌کارلو (MCTS) برای خودروهای خودران برای انجام مانورهای مختلف به‌منظور جلوگیری از برخورد پیشنهاد کرد. آن‌ها فرآیند کنترل را به‌عنوان یک مشکل فرایند مارکوف مدل کردند و از MCTS برای تولید زاویه فرمان استفاده کردند. پیشنهاد آن‌ها مقاومت بیشتر کنترل و موفقیت بالاتر در اجتناب از رویدادهای غیرمنتظره را نشان داد. همچنین [Artunedo et al, 2024] یک ارزیابی مقایسه‌ای از روش‌های کنترل جانبی مختلف برای وسایل نقلیه خودران ارائه می‌دهد. راهبردهای کنترلی مانند تنظیم‌کننده درجه دوم خطی LQR، کنترل بدون مدل MFC، کنترل بدون مدل با سرعت تطبیقی SAMFC، PID و کنترل پیش‌بینی مدل غیرخطی NLMPCرا ارزیابی می‌کند. این مطالعه همچنین عملکرد این روش‌ها را در سناریوهای مختلف رانندگی با در نظر گرفتن عواملی مانند ثبات وسیله نقلیه، راحتی سرنشینان و نرمی کنترل ارزیابی می‌کند. اما، تمرکز این تحقیقات بیشتر بر عملکرد نهایی الگوریتم‌ها بوده است تا تأثیر ساختار کنترلی (تعداد و نقش عامل[۱۸]ها) بر فرآیند آموزش.

با وجود پیشرفت‌های چشمگیر در زمینه کنترل خودروهای خودران، در هیچ یک از پژوهش‌های پیشین، مقایسه مستقیم و جامعی بین یک سیستم کنترل تنها با یک عامل و یک سیستم با عامل‌های مجزا برای کنترل طولی و عرضی انجام نشده است و تأثیر این تفکیک بر سرعت بالاتر آموزش شبکه و دقت بیشتر آن به درستی مورد بررسی قرار نگرفته است. این مقاله با ارائه این مقایسه، به این شکاف تحقیقاتی می‌پردازد.

از نوآوری‌های این پژوهش می‌توان به (۱) تولید ورودی‌های کنترل ترکیبی شتاب طولی و جانبی خودروی بدون سرنشین با استفاده از شبکه عصبی عمیق به صوت هم‌زمان، (۲) ارائه تابع پاداش مناسب جهت آموزش صحیح شبکه عصبی عمیق و (۳) ارائه الگوریتم جدید کنترلی جداسازی عامل‌های کنترل شتاب طولی و عرضی خودروی بدون سرنشین در یادگیری تقویتی عمیق اشاره کرد. جدول ۱ دسته‌بندی کلی‌ای از مطالعات پیشین، شامل روش انتخاب شده و تمرکز الگوریتم کنترلی ارائه می‌دهد.

جدول ۱. جمع‌بندی و دسته‌بندی ادبیات پژوهش

مرجع	نوع کنترل	روش استفاده شده	عملکرد
[Mao et al, 2024]	ترکیبی	RL سایر روش‌های و DDPG	یک مرور جامع از یادگیری عمیق در خودروهای خودران که به طور کلی به روش‌های DRL می‌پردازد.
[Ning et al, 2021]	ترکیبی	معماری‌های AI	بررسی معماری‌های موجود AI و معرفی مفهوم AI انسانی
[Liu and Diao, 2024]	ترکیبی	Q-learning، PPO ،DDPG	یک مرور کلی بر الگوریتم‌های کنترل بهینه با استفاده از یادگیری تقویتی
[Li et al, 2023]	ترکیبی	DDPG، PPO ،MADDPG	مروری بر یادگیری تقویتی برای برنامه‌ریزی حرکت در خودروهای خودران با تمرکز بر این الگوریتم‌ها.
[Guo et al, 2020]	عرضی	DDPG، MADDPG	یک راهبرد کنترل جانبی برای تغییر لاین، که باعث بهبود جریان ترافیک می‌شود.
[Wu et al, 2020]	طولی	Q-learning	یک روش برای کنترل سرعت با استفاده از DVSL، که ایمنی در بزرگراه را افزایش می‌دهد.
[Zhang et al, 2018]	طولی	DQN، Double Q-learning	رویکردی برای کنترل سرعت بر اساس DQN و Double Q-learning
[Van et al, 2016]	ترکیبی	Double Q-learning	یک مقاله پایه‌ای که الگوریتم Double Q-learning را معرفی می‌کند.
[Baheri et al, 2020]	عرضی	DQN، DDAC	روشی برای حفظ لاین در رانندگی شهری که نتایج موفقی در شرایط مختلف آب‌وهوایی دارد.
[Ye et al, 2020]	عرضی	PPO	راهبرد تغییر لاین خودکار در بزرگراه‌ها که مانورهای نرم و ایمن را می‌آموزد.
[Toromanoff et al, 2020]	ترکیبی	DQN	رویکردی برای حل شرایط پیچیده رانندگی شهری که قابل تعمیم به محیط‌های ناشناخته است.
[Pérez-Gil et al, 2022]	ترکیبی	DQN vs DDPG	مقایسه مستقیم DQN و DDPG که نشان می‌دهد DDPG عملکرد بهتری دارد.
[Lee et al, 2022]	ترکیبی	IRL	الگوریتم جدیدی برای یادگیری تقویتی معکوس که در حفظ لاین و تغییر لاین ارزیابی می‌شود.
[Du et al, 2023]	طولی	EK-DDPG	بهبود قابل توجه راحتی و راندمان محاسباتی در کنترل تعلیق نسبت به DDPG و MPC
[Chen et al, 2020]	ترکیبی	MCTS-based RL	افزایش مقاومت و موفقیت کنترل در مانورهای اجتناب از برخورد با استفاده از جستجوی درخت مونت‌کارلو
[Artunedo et al, 2024]	عرضی	LQR، MFC ،PID و NLMPC	یک ارزیابی مقایسه‌ای از روش‌های کنترل کلاسیک، با تمرکز بر ثبات، راحتی و نرمی کنترل

۳. مدل دینامیکی خودرو

مدل دینامیکی دقت بیشتری نسبت به مدل سینماتیک دارد به این معنا که شامل نیروهای وارد بر خودرو و به‌ویژه نیروهای تایر می‌شود. قانون دوم حرکت نیوتن و روش‌های اویلر لاگرانژ بر روی سیستم خودرو برای به دست آوردن مدل دینامیکی اعمال می‌شوند. مدل دینامیکی کامل بسیار پیچیده و غیرخطی است که برای حرکات انتقالی و چرخشی در فضای سه‌بعدی، برای یک وسیله نقلیه کامل با چهارچرخ محاسبه می‌شود. چنین مدل‌هایی عمدتاً برای اهداف اعتبارسنجی استفاده می‌شوند و برای طراحی کنترل‌کننده بسیار پیچیده هستند. به‌جای آن از مدل‌های دینامیکی دوچرخ ساده‌شده استفاده می‌شود.

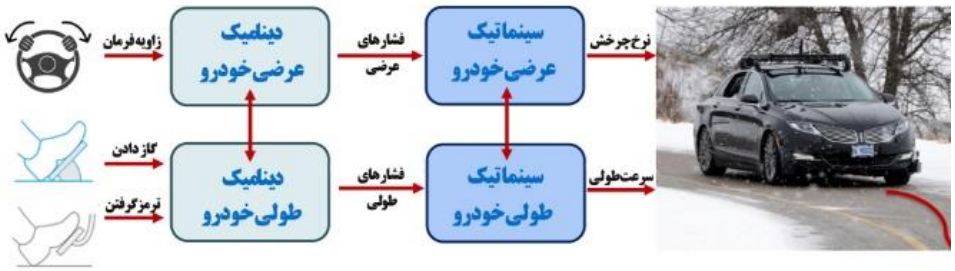
مدل دینامیکی دوچرخه برای حرکت مسطح دوبعدی استفاده می‌شود. انتقال در محور x وy طولی/جانبی و چرخش حول محور z، یک مدل وسیله نقلیه ۳ درجه آزادی [۱۹] را ایجاد می‌کند و در برخی موارد، دینامیک طولی نادیده گرفته می‌شود، مانند ردیابی مسیر که در آن وظیفه به کنترل دینامیک جانبی و حرکت انحرافی کاهش می‌یابد که منجر به مدل ۲ درجه آزادی می‌شود. با در نظر گرفتن مدل دوچرخه ساده و اعمال قوانین نیوتن معادلات مدل زیر حاصل می‌شود [Huang et al ,2022]:

$$\begin{aligned}\sum F_x &= ma_x \\ \sum F_y &= ma_y \\ \sum M_z &= I_z\ddot{\psi}\end{aligned} \qquad (۱)$$

که در آن I_z ممان اینرسی، M_z گشتاور چرخشی حول محور z و F_x و F_y به ترتیب نیروهای طولی و جانبی هستند. این نیروها روی چرخ‌های جلو و عقب اعمال می‌شوند. a_x و a_y شتاب‌های اینرسی طولی و جانبی هستند و برحسب شتاب‌های طولی/جانبی، نرخ انحراف بیان می‌شوند.

$$\begin{aligned}a_x &= \ddot{x} - \dot{\psi}\dot{y} \\ a_y &= \ddot{y} - \dot{\psi}\dot{x}\end{aligned} \qquad (۲)$$

در سرعت‌های بالاتر خودرو، به‌جای یک مدل سینماتیک، باید یک مدل دینامیکی برای حرکت جانبی وسیله نقلیه ایجاد شود.



شکل ۱. نحوه کنترل فاز طولی و عرضی خودروی بدون سرنشین

۳-۱ دینامیک عرضی خودرو

با نادیده گرفتن زاویه کرانه جاده و اعمال قانون دوم نیوتن برای حرکت در امتداد محور y داریم:

$$ma_y = F_{yf} + F_{yr} \qquad (۳)$$

که $a_y = \left(\frac{d^2y}{dt^2}\right)$ شتاب اینرسی خودرو در مرکز ثقل خودرو، در جهت محور y و F_{yf} و F_{yr} به ترتیب نیروهای جانبی لاستیک چرخ‌های جلو و عقب هستند. دو عبارت به a_y کمک می‌کنند که یکی شتاب $\ddot{y}$ که ناشی از حرکت در امتداد محور yاست و دیگری شتاب مرکز محور $V_x\dot{\psi}$، ازاین‌رو:

$$a_y = \ddot{y} + V_x\dot{\psi} \qquad (۴)$$

در ادامه معادله حرکت انتقالی جانبی وسیله نقلیه به‌صورت معادله (۵) به‌دست‌آمده است.

$$m(\ddot{y} + V_x\dot{\psi}) = F_{yf} + F_{yr} \qquad (۵)$$

با اعمال ممان اینرسی حول محور z معادله دینامیک انحراف به دست می‌آید.

$$I_z\ddot{\psi} = \ell_f F_{yf} + \ell_r F_{yr} \qquad (۶)$$

گام بعدی مدل‌سازی نیروهای جانبی تایر F_{yf} و F_{yr} است که بر روی وسیله نقلیه اعمال می‌شود. نتایج تجربی نشان می‌دهد که نیروی جانبی تایر متناسب با «زاویه لغزش» برای زوایای لغزش کوچک است. زاویه لغزش یک تایر به‌عنوان زاویه بین جهت لاستیک و جهت بردار سرعت چرخ تعریف می‌شود.

$$\alpha_f = \delta - \theta_{vf} \qquad (۷)$$

که θ_{vf} زاویه‌ای است که بردار سرعت با محور طولی وسیله نقلیه ایجاد می‌کند و δ زاویه فرمان چرخ جلو است. زاویه لغزش عقب نیز به‌طور مشابه به دست خواهد آمد.

$$\alpha_r = -\theta_{vr} \qquad (۸)$$

بنابراین نیروی جانبی تایر برای چرخ‌های جلوی خودرو را می‌توان به‌صورت زیر نوشت:

$$F_{yf} = 2C_{\alpha_f}(\delta - \theta_{vf}) \qquad (۹)$$

که در آن ثابت تناسب C_{α_f} ، سفتی هر تایر جلو نامیده می‌شود، δزاویه فرمان چرخ جلو و θ_{vf} زاویه سرعت تایر جلو است. عامل ۲ به دلیل این است که دوچرخ جلو وجود دارد. به‌طور مشابه، لاستیک جانبی برای چرخ‌های عقب را می‌توان به‌صورت زیر نوشت:

$$F_{yr} = 2C_{\alpha_r}(-\theta_{vr}) \qquad (۱۰)$$

که در آن C_{α_r} سفتی هر تایر عقب و θ_{vr} زاویه سرعت لاستیک عقب است. برای محاسبه θ_{vf} و θ_{vr} می‌توان از روابط زیر استفاده کرد.

$$\tan(\theta_{vf}) = \frac{V_y + \ell_f \dot{\psi}}{V_x}$$
$$\tan(\theta_{vr}) = \frac{V_y - \ell_r \dot{\psi}}{V_x} \qquad (۱۱)$$

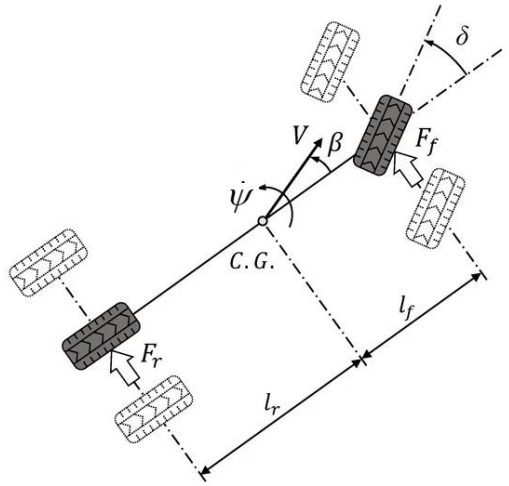
با فرض $V_y = \dot{y}$ و کوچک بودن زاویه انحراف داریم:

$$\theta_{vf} = \frac{\dot{y} + \ell_f \dot{\psi}}{V_x}$$
$$\theta_{vr} = \frac{\dot{y} + \ell_r \dot{\psi}}{V_x} \qquad (۱۲)$$

با جایگزینی، مدل فضای حالت را می‌توان به‌صورت زیر نوشت:

$$\frac{d}{dt}\begin{Bmatrix} y \\ \dot{y} \\ \psi \\ \dot{\psi} \end{Bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{2C_{\alpha_f}+2C_{\alpha_r}}{mV_x} & 0 & -V_x - \frac{2C_{\alpha_f}+2C_{\alpha_r}}{mV_x} \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{2\ell_f C_{\alpha_f} - 2\ell_r C_{\alpha_r}}{I_z V_x} & 0 & -\frac{2\ell_f^2 C_{\alpha_f} - 2\ell_r^2 C_{\alpha_r}}{I_z V_x} \end{bmatrix} \begin{bmatrix} y \\ \dot{y} \\ \psi \\ \dot{\psi} \end{bmatrix} + \begin{Bmatrix} 0 \\ \frac{2C_{\alpha_f}}{m} \\ 0 \\ \frac{2\ell_f C_{\alpha_f}}{I_z} \end{Bmatrix} \delta \qquad (۱۳)$$

هنگامی‌که هدف توسعه یک سیستم کنترل فرمان برای حفظ خطوط خودکار است، استفاده از یک مدل پویا که در آن متغیرهای حالت، خطای موقعیت و خطای جهت‌گیری نسبت به جاده هستند، مفید است؛ بنابراین مدل جانبی توسعه‌یافته برحسب متغیرهای خطای e_1، فاصله مرکز ثقل وسیله نقلیه از خط مرکزی لاین و e_2، خطای جهت‌گیری وسیله نقلیه نسبت به سمت جاده دوباره تعریف می‌شود.



شکل ۲. مدل دینامیکی دوچرخه در حرکت عرضی [Cao. et al. 2023]

وسیله نقلیه‌ای را در نظر بگیرید که با سرعت طولی ثابت V در جاده‌ای با شعاع ثابت R حرکت می‌کند. مجدداً فرض کنید که شعاع R بزرگ است تا بتوان همان مفروضات زاویه کوچکی را که در بخش قبل انجام داد انجام داد. نرخ تغییر جهت مورد نظر وسیله نقلیه به‌صورت تعریف می‌شود:

$$\dot{\psi}_{des} = \frac{V_x}{R} \qquad (۱۴)$$

سپس شتاب مورد نظر وسیله نقلیه را می‌توان به‌صورت زیر نوشت:

$$\frac{V_x^{\,2}}{R} = V_x \dot{\psi}_{des} \qquad (۱۵)$$

سپس

$$\begin{aligned} \ddot{e}_1 &= (\ddot{y} + V_x\dot{\psi}) - \frac{V_x}{R} = \ddot{y} + V_x(\dot{\psi} - \dot{\psi}_{des}) \\ e_2 &= (\psi - \psi_{des}) \\ \dot{e}_1 &= \dot{y} + V_x(\psi - \psi_{des}) \end{aligned} \quad (۱۶)$$

این معادلات با فرض ثابت بودن سرعت V_x صحیح است.

$$m\ddot{e}_1 = \dot{e}_1\left[-\frac{2}{V_x}C_{\alpha_f} - \frac{2}{V_x}C_{\alpha_r}\right] + e_2\left[2C_{\alpha_f} + 2C_{\alpha_r}\right] + \dot{e}_2\left[-\frac{2C_{\alpha_f}\ell_f}{V_x} + \frac{2C_{\alpha_r}\ell_r}{V_x}\right] + \dot{\psi}_{des}\left[-\frac{2C_{\alpha_f}\ell_f}{V_x} + \frac{2C_{\alpha_r}\ell_r}{V_x}\right] + 2C_{\alpha_f}\delta \quad (۱۷)$$

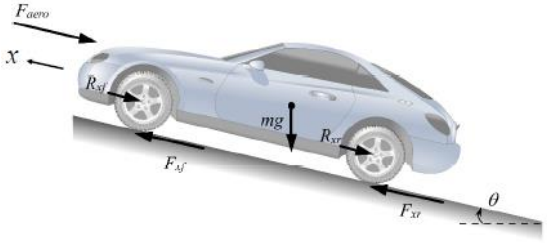
$$I_z\ddot{e}_2 = 2C_{\alpha_f}\ell_f\delta + \dot{e}_1\left[-\frac{2C_{\alpha_f}\ell_f}{V_x} + \frac{2C_{\alpha_r}\ell_r}{V_x}\right] + e_2\left[2C_{\alpha_f}\ell_f - 2C_{\alpha_r}\ell_r\right] + \dot{e}_2\left[-\frac{2C_{\alpha_f}\ell_f^2}{V_x} - \frac{2C_{\alpha_r}\ell_r^2}{V_x}\right] - I_z\ddot{\psi}_{des} + \dot{\psi}_{des}\left[-\frac{2C_{\alpha_f}\ell_f^2}{V_x} - \frac{2C_{\alpha_r}\ell_r^2}{V_x}\right] \quad (۱۸)$$

بنابراین مدل فضای حالت در متغیرهای خطای ردیابی با استفاده از معادله زیر به دست خواهد آمد:

$$\frac{d}{dt}\begin{Bmatrix} e_1 \\ \dot{e}_1 \\ e_2 \\ \dot{e}_2 \end{Bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{2C_{\alpha_f}+2C_{\alpha_r}}{mV_x} & \frac{2C_{\alpha_f}+2C_{\alpha_r}}{m} & -V_x - \frac{2C_{\alpha_f}+2C_{\alpha_r}}{mV_x} \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{2\ell_f C_{\alpha_f} - 2\ell_r C_{\alpha_r}}{I_z V_x} & -\frac{2C_{\alpha_f}\ell_f - 2C_{\alpha_r}\ell_r}{I_z} & -\frac{2\ell_f^2 C_{\alpha_f} - 2\ell_r^2 C_{\alpha_r}}{I_z V_x} \end{bmatrix} \begin{bmatrix} e_1 \\ \dot{e}_1 \\ e_2 \\ \dot{e}_2 \end{bmatrix} + \begin{Bmatrix} 0 \\ \frac{2C_{\alpha_f}}{m} \\ 0 \\ \frac{2\ell_f C_{\alpha_f}}{I_z} \end{Bmatrix}\delta + \begin{bmatrix} 0 \\ -\frac{2C_{\alpha_f}\ell_f - 2C_{\alpha_r}\ell_r}{mV_x} - V_x \\ 0 \\ -\frac{2\ell_f^2 C_{\alpha_f} - 2\ell_r^2 C_{\alpha_r}}{I_z V_x} \end{bmatrix}\dot{\psi}_{des} = AX + B_1\delta + B_2\dot{\psi}_{des} \quad (۱۹)$$

۳-۲ دینامیک طولی خودرو

در مدل دینامیک طولی، خودرو تحت تأثیر نیروهای طولی تایر، نیروهای کشش آئرودینامیکی، نیروهای مقاومت غلتشی و نیروهای گرانشی است. سیستم انتقال قدرت طولی خودرو شامل موتور احتراق داخلی، مبدل گشتاور، گیربکس و چرخ‌ها است. همان‌طور که در شکل زیر نشان داده شده است، خودرویی را در نظر بگیرید که در یک جاده شیب‌دار حرکت می‌کند. نیروهای طولی خارجی وارد بر خودرو شامل نیروهای کششی آئرودینامیکی، نیروهای گرانشی، نیروهای طولی تایر و نیروهای مقاومت غلتشی است.



شکل ۳. نیروهای دینامیکی در حرکت طولی خودرو

[Rajamani, 2006]

با توجه به نیروهای وارد بر خودرو در حرکت طولی مطابق شکل ۳ رابطه قانون دوم نیوتن برای آن به‌صورت زیر نوشته می‌شود:

$$m\ddot{x} = F_{xf} + F_{xr} - F_{aero} - R_{xf} - R_{xr} - mg\sin(\theta) \quad (۲۰)$$

که در آن F_{xf} و F_{xr} به ترتیب نیروی طولی تایر در لاستیک‌ها جلو و عقب هستند. F_{aero} نیروی کشش آئرودینامیکی طولی معادل، R_{xf} و R_{xr} نیروی ناشی از مقاومت غلتشی در تایرهای جلو، m جرم وسیله نقلیه، g شتاب ناشی از گرانش و θ زاویه شیب جاده‌ای است که وسیله نقلیه در آن حرکت می‌کند. هنگامی‌که جهت طولی حرکت x به سمت چپ باشد، زاویه θ در جهت عقربه‌های ساعت، مثبت تعریف می‌شود و هنگامی‌که جهت طولی حرکت x به سمت راست باشد، در خلاف جهت عقربه‌های ساعت مثبت تعریف می‌شود[Huang et al, 2022].

فشار طولی چرخ جلو $F_{xf} = C_{\sigma f}\sigma_{xf}$ که $\sigma_{xf} = \frac{r_{eff}\omega_{wf} - \dot{x}}{\dot{x}}$ در زمان ترمز و $\sigma_{xf} = \frac{r_{eff}\omega_{wf} - \dot{x}}{r_{eff}\omega_{wf}}$ در زمان شتاب‌گیری

فشار طولی چرخ عقب $F_{xr} = C_{\sigma r}\sigma_{xr}$ که $\sigma_{xr} = \frac{r_{eff}\omega_{wr} - \dot{x}}{\dot{x}}$ در زمان ترمز و $\sigma_{xr} = \frac{r_{eff}\omega_{wr} - \dot{x}}{r_{eff}\omega_{wr}}$ در زمان شتاب‌گیری

مقاومت چرخش لاستیک‌ها $R_{xf} + R_{xr} = f(F_{zf} + F_{zr})$ که در آن مطابق زیر است:

$$\begin{aligned} F_{zf} &= \frac{-F_{aero}h_{aero} - m\ddot{x}h + mgh\sin(\theta) + mg\ell_r\cos(\theta)}{\ell_f + \ell_r} \\ F_{zr} &= \frac{-F_{aero}h_{aero} - m\ddot{x}h + mgh\sin(\theta) + mg\ell_f\cos(\theta)}{\ell_f + \ell_r} \end{aligned} \quad (۲۱)$$

و فشار درگ آئرودینامیک برابر است با

$$F_{aero} = \frac{1}{2}\rho C_d A_F (\dot{x} + V_{wind})^2 .$$

۴. استفاده از شبکه عصبی در کنترل

یادگیری تقویتی زیرشاخه‌ای از یادگیری ماشینی است که به مسئله یادگیری خودکار و تصمیم‌گیری‌های بهینه در طول زمان می‌پردازد. درحالی‌که روش‌های یادگیری عمیق بر توسعه برنامه‌هایی تمرکز می‌کنند که از داده‌ها برای یادگیری به‌طور مستقل استفاده می‌کنند، روش‌های یادگیری تقویتی به یک عامل هوشمند اجازه می‌دهد تا از خطاها و تجربیات خود بیاموزد و در جهت بهبود عملکرد، از آن‌ها استفاده کند. عامل یادگیری تقویتی با فعالیت در محیط، پاداش دریافت می‌کند. هدف آن انتخاب عملی است که پاداش تجمعی مورد انتظار را در طول زمان به حداکثر برساند [Wang et al, 2024].

یکی از چالش‌های اصلی در یادگیری تقویتی، مدیریت مبادله بین اکتشاف و بهره‌برداری است. برای به حداکثر رساندن پاداش‌هایی که دریافت می‌کند، یک عامل باید با انتخاب اقداماتی که منجر به پاداش بالا می‌شوند از دانش خود بهره‌برداری کند. از سوی دیگر، برای کشف چنین اقدامات سودمندی، باید ریسک انجام اقدامات جدیدی را بپذیرد که ممکن است به پاداش‌های بالاتر یا پایین‌تری نسبت به اقدامات با ارزش فعلی برای هر وضعیت سیستم منجر شود؛ به‌عبارت‌دیگر، عامل یادگیری باید برای به دست آوردن پاداش بیشتر تلاش کند و برای این کار باید ناشناخته‌ها را کشف کند تا در آینده اقدام‌های بهتری انجام دهد. همچنین، عامل باید در ابتدای فرآیند آموزش، زمانی که اطلاعات کمی در مورد محیط مسئله وجود دارد، بیشتر کاوش کند و شناخت بیشتری از محیط به دست آورد. البته به این نکته توجه شود که طراحی راهبردهای اکتشاف محیط، برای عوامل یادگیری تقویتی هنوز یک حوزه باز در تحقیقات است (به‌عنوان مثال [Bellotti et al,2023]).

۴-۱ تعریف و مبانی یادگیری تقویتی

برنامه‌ریزی پویا (DP) [۲۰] به مجموعه‌ای از الگوریتم‌ها اطلاق می‌شود که می‌توانند برای محاسبه سیاست‌های بهینه با توجه به یک مدل کامل از محیط ازنظر توابع پاداش استفاده شوند. برخلاف DP، در روش‌های مونت‌کارلو هیچ فرضی بر دانش کامل محیطی وجود ندارد از سوی دیگر، روش‌های تفاوت زمانی [۲۱] (TDL)، می‌تواند برای سناریوهای غیر اپیزودیک قابل استفاده باشد؛ یک عامل یادگیری تقویتی را می‌توان به‌عنوان یک فرآیند تصمیم‌گیری مارکوف (MDP) [۲۲] به‌صورت عاملی که با اجرای اقدامات و دریافت مشاهدات و پاداش‌ها با محیط تعامل دارد، تعریف کرد. قبل از توضیح بیشتر لازم است تا تعاریف زیر بیان شود.

وضعیت [۲۳] (State|S): شرایطی از محیط است که عامل در هرلحظه با آن مواجه می‌شود و با انجام اقداماتی بر محیط، این پارامتر را تغییر می‌دهد. به بیانی دیگر، وضعیت هرگونه اطلاعاتی از محیط است که در هر مرحله زمانی به عامل داده می‌شود تا به کمک آن بتواند اقدام مناسبی را انجام دهد. اگر اطلاعات ارائه‌شده توسط وضعیت به‌گونه‌ای باشد که بتوان به کمک آن وضعیت‌های آینده محیط را با توجه به اقدامات عامل، تعیین نمود، در آن صورت گفته می‌شود که وضعیت دارای ویژگی مارکوفی است.

اقدام [۲۴] (Action|A): شامل تمامی واکنش‌های محتملی است که عامل تصمیم‌گیرنده ممکن است در مواجهه با وضعیت ایجادشده، از خود نشان دهد. اقدام‌ها با توجه به متناهی بودن و یا نامتناهی بودنشان به دو نوع اپیزودیک و یا مستر (غیر اپیزودیک) تقسیم‌بندی می‌شوند. منظور از اقدامات اپیزودیک، وجود یک نقطه آغاز و یا پایان برای آن‌هاست. ولی در وظایف مستمر یا غیر اپیزودیک نقطه پایانی برای اقدامات وجود ندارد. در این زمینه عامل تا زمانی که توسط نیروی خارجی (مثل کاربر انسانی) متوقف نشود، همواره به یادگیری و انجام اقدام مناسب به تعامل با محیط می‌پردازد.

پاداش [۲۵] (Reward|R): بازخوردی که پس از ارزیابی هر اقدام عامل، توسط محیط برای آن ارسال می‌شود را پاداش می‌گویند.

سیاست [۲۶] (Policy|π): عامل تصمیم‌گیرنده برای انجام هر عملی بر روی محیط و پاسخ به وضعیت فعلی، راهبردی در پیش می‌گیرد که به آن در اصطلاح سیاست گفته می‌شود.

ارزش [۲۷] (Value|V): ارزش عبارت است از سود مورد انتظار در بلندمدت که ناشی از وضعیت کنونی s تحت سیاست π است. به بیانی دیگر، ارزش، پاداش بلندمدت مورد انتظار تنزیل‌شده‌ای است که برای بلندمدت اعمال می‌شود.

تابع انتقال: مقدار توزیع احتمال رفتن به وضعیت جدید با توجه به حالت اولیه محیط و عملکرد اعمال‌شده بر آن را، مشخص می‌کند.

۲-۴ فرآیند مارکوف

فرآیندهای تصمیم مارکوف در مسائل تصمیم‌گیری تک سیاستِ، به‌عنوان یک روش بسیار مناسب در نظر گرفته می‌شوند. معمولاً یک فرایند تصمیم‌گیری مارکوف به‌صورت چندتایی <S,A,P,R>داده می‌شود. متغیر S مجموعه‌ای از همه حالت‌های ممکن است. متغیر A مجموعه‌ای از فعالیت‌هایی است که برای هر عامل در دسترس است. متغیر P تابع گذار است. به‌عنوان ورودی، حالت فعلی، عملی که روی آن انجام‌شده و حالت بعدی را دریافت می‌کند و احتمال انتقال به حالت بعدی پیشنهادشده را به‌عنوان خروجی برمی‌گرداند. توجه شود که P به S_t بستگی دارد و به S_{t-1} وابسته نیست.

$P_a(s,s') = P(s_{t+1} = s' | s_t = s, a_t = a)$ احتمال اینکه اقدام a در حالت s و در زمان t منجر به حالت s' در زمان $t+1$ شود را نشان می‌دهد.

متغیر R تابع تشویق (جایزه) است. این تابع به‌عنوان ورودی، حالت فعلی و حرکتی که انجام‌شده را گرفته و امتیازی که عامل با انجام دادن آن حرکت به دست می‌آورد را نمایش می‌دهد و معمولاً همه مقدارهای گویا را می‌پذیرد و به‌صورت زیر تعریف می‌شود:

$$R : s_t \in S, a \in A \rightarrow R \qquad (۲۲)$$

سیاست تصادفی $\pi: S \rightarrow D$ یک نگاشت از فضای حالت به یک احتمال در مجموعه اقدامات است و $\pi(a|s)$ احتمال انتخاب عمل a در حالت s را نشان می‌دهد. هدف یافتن سیاست بهینه π^*است [Lamssaggad et al, 2021]

$$\pi^* = \arg\max_{\pi} \underbrace{E_\pi \left\{ \sum_{k=0}^{H-1} \gamma^k r_{k+1} | s_0 = s, a_0 = a \right\}}_{=V_\pi(s)} \qquad (۲۳)$$

برای همه حالات $s \in S$ ، $r_k = R(s_k, a_k)$ پاداش در زمان k است و $V_\pi(s)$ ، تابع ارزش در وضعیت s به دنبال سیاستπ ، است که با اضافه شدن اکشن a معادله Q به‌صورت زیر به دست می‌آید:

$$Q_\pi(s,a) = E_\pi \left\{ \sum_{k=0}^{H-1} \gamma^k r_{k+1} | s_0 = s, a_0 = a \right\} \qquad (۲۴)$$

ضریب فراموشی $\gamma \in [0,1]$نحوه توجه یک عامل به پاداش‌های آینده را کنترل می‌کند. مقادیر پایین γ برای جایی که هدف یک عامل به حداکثر رساندن پاداش‌های کوتاه‌مدت است مناسب است، درحالی‌که مقادیر بالای γ باعث می‌شود که عوامل آینده‌نگرتر باشند و پاداش‌ها را در بازه زمانی طولانی‌تری به حداکثر برسانند. افق H به مقدار افق زمانی در فرایند مارکوف اشاره دارد. در مسائل افق نامتناهی $H = \infty$ است، درحالی‌که در حوزه‌های اپیزودیک H مقدار محدودی دارد. آخرین حالتی که در یک حوزه اپیزودیک به دست می‌آید، حالت پایانی نامیده می‌شود. در حوزه‌های افق محدود یا هدف‌محور، عوامل γ کوچک‌تر ممکن است برای تشویق عوامل به تمرکز بر روی دستیابی به هدف، مورد استفاده قرار گیرند، درحالی‌که در حوزه‌های افق نامتناهی ممکن است از عوامل γ بزرگ‌تر برای رسیدن به هدف استفاده شود. یک فرایند مارکوف ویژگی مارکوف را برآورده می‌کند، یعنی انتقال وضعیت سیستم فقط به آخرین وضعیت و عمل وابسته است، نه به تاریخچه کامل حالت‌ها و اقدامات در فرآیند تصمیم‌گیری.

لازم به ذکر است که در برخی از موارد یافتن وضعیت‌هایی از محیط که خاصیت مارکوفی داشته باشد، کار دشواری است. خبر

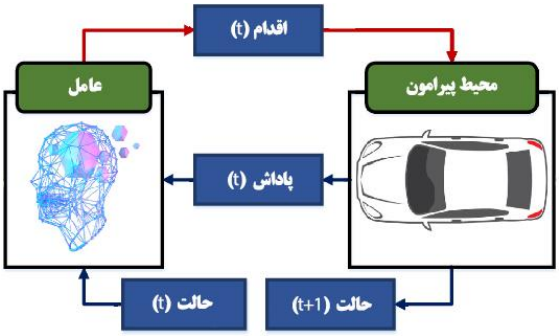
خوب این است، الگوریتم‌های یادگیری تقویتی بر روی تقریب‌هایی با ویژگی‌هایی نزدیک به مارکوف هم توانسته عملکرد خوبی از خود نشان دهد [Mahaadevan et al,2024]. در چنین مواقعی، فرض می‌شود که وضعیتی از محیط با ویژگی مارکوفی وجود دارد که ما نمی‌توانیم آن را مشاهده کنیم. آنچه ما مشاهده می‌کنیم تنها بخشی از این وضعیت است که به آن مشاهده جزئی می‌گویند و به فرایند حل آن فرآیند تصمیم‌گیری مارکوف با مشاهده‌پذیری جزئی [۲۸] می‌گویند.

اکثر مدل‌های بدون مدل از دو رویکرد ارزش محور و یا سیاست محور استفاده می‌کنند. در رویکرد سیاست محور، هدف بهینه‌سازی تابع سیاست است بدون اینکه به تابع ارزش کار داشته باشیم. به بیانی دیگر عامل یک تابع سیاست را می‌آموزد، آن را در حین یادگیری در حافظه نگه می‌دارد و سعی می‌کند هر وضعیت را به بهترین اقدام ممکن نگاشت کند. لازم به ذکر است سیاست‌ها ممکن است قطعی (برای یک وضعیت، همیشه اقدام مشابهی را بازمی‌گرداند) و یا تصادفی (برای هر اقدام یک توزیع احتمالی در نظر می‌گیرد) باشند.

در رویکرد ارزش محور، برخلاف رویکرد سیاست‌محور که به تابع ارزش کاری ندارد، هدف بهینه‌سازی تابع ارزش خواهد بود. به‌عبارت‌دیگر، عامل اقدامی را انتخاب می‌نماید که برآورد می‌کند بیشترین پاداش را در آینده دریافت خواهد کرد.

۴-۳ کاربردهای هوش مصنوعی در صنعت خودرو

یک وسیله نقلیه خودران می‌تواند محیط خود را تشخیص دهد و بدون دخالت انسان تصمیم‌گیری کند [Yuan, Shan and Mi, 2023]. خودروهای خودران با تکیه‌بر ارتباطات خودرو، پیام‌های ایمنی، شرایط ترافیکی و پیام‌های هشداردهنده در صورت ترافیک یا تصادف را مبادله می‌کنند. وسایل نقلیه بدون راننده در درجه اول به مجموعه‌ای از حسگرها، محرک‌ها، الگوریتم‌های پیچیده، روش‌های هوش مصنوعی و منابع محاسباتی قدرتمند برای اجرای نرم‌افزار متکی هستند. به‌این‌ترتیب، خودروهای خودران قادر به مقابله با موقعیت‌های جاده‌ای پیچیده هستند که منجر به بهبود قابل توجه ایمنی، راحتی و راحتی کاربر می‌شود.



شکل ۴. ساختار یادگیری تقویتی در کنترل خودروی بدون سرنشین

برای امکان‌پذیر ساختن یک وسیله نقلیه سطح ۵، لازم است که وسیله نقلیه مانند یک راننده انسان، «فکر»، «درک» و «واکنش» داشته باشد. دستاوردهای اخیر هوش مصنوعی در زمینه‌های مختلف، به‌ویژه در طبقه‌بندی تصویر، تشخیص اشیا و تشخیص گفتار منجر به استفاده روزافزون از روش‌های هوش مصنوعی مانند یادگیری عمیق و یادگیری تقویتی برای تحقق وسایل نقلیه سطح ۵ شده است. رویکردهای مبتنی بر یادگیری عمیق مطالعات بسیاری را برای مقابله با مسائل چالش‌برانگیز مختلف در خودروهای خودران، مانند تشخیص دقیق و مکان‌یابی موانع در جاده‌ها، کنترل مناسب خودرو و برنامه‌ریزی حرکت، قادر ساخته است. این پژوهش صرفاً بر روی رویکردهای مبتنی بر یادگیری عمیق و یادگیری تقویتی متمرکز است، بنابراین گنجاندن روش‌های مبتنی بر یادگیری ماشین کم‌عمق [۲۹] (SML) را حذف می‌کند، موضوعی که درگذشته به‌طور گسترده مورد بررسی قرارگرفته است.

روش‌های یادگیری عمیق و یادگیری تقویتی به دلیل توانایی‌شان در بهبود وظایف خودروهای خودران معروف هستند. برای مثال، یادگیری عمیق نتایج امیدوارکننده‌ای را در تشخیص اشیا نشان داد که آن را برای درک محیط در رانندگی خودکار مناسب می‌کند. یادگیری تقویتی با موفقیت در زمینه‌های دیگری مانند بازی و رباتیک استفاده شده است، جایی که نیاز به یادگیری از

محیط وجود دارد. اخیراً، روش‌های یادگیری تقویتی درنتیجه بسیاری از نتایج امیدوارکننده، مورد توجه جامعه تحقیقاتی خودروهای خودران قرارگرفته است.

۵. روش پژوهش

این پژوهش به‌منظور بررسی و مقایسه دو ساختار مختلف کنترل خودروی بدون سرنشین با استفاده از یادگیری تقویتی انجام شد. تمرکز اصلی بر روی مقایسه عملکرد عامل یادگیری تقویتی یکسان و چند عامل مستقل جداگانه برای کنترل طولی و عرضی خودرو است.

در حالت عامل یکسان، یک عامل واحد، وظیفه یادگیری و تولید هر دو فرمان طولی (مانند شتاب یا ترمز) و عرضی (مانند زاویه فرمان) را به طور هم‌زمان بر عهده دارد. در حالت عامل های جدا، دو عامل مستقل به‌صورت مجزا، اما با در نظر گرفتن تاثیر بروی یکدیگر آموزش می‌بینند و هرکدام فقط مسئول یکی از جنبه‌های کنترلی (کانال طولی و عرضی) هستند. در نتیجه هر عامل فقط بخشی از اطلاعات مربوط به زیرسیستم خودش را به عنوان ورودی دریافت می‌کند وخروجی تولیدی تنها مربوط به آن کانال است.

ابتدا یک سناریوی رانندگی استاندارد با جاده منحنی (شعاع ثابت) و حضور خودروی جلویی به‌عنوان مرجع فاصله تعریف شد. وظیفه عامل هوشمند حفظ موقعیت در لاین و رعایت فاصله ایمن با خودروی جلو است. دلیل اینکه جاده منحنی در نظر گرفته شده است این است که روند کنترل هم در کانال طولی و هم کانال عرضی بررسی شود.

در ساختار اول یا یک عامل یکسان، طراحی با ورودی ترکیبی شامل اطلاعات عرضی و طولی و خروجی عامل شامل دو مقدار زاویه فرمان و شتاب طولی حرکت است. آموزش عامل در این حالت با استفاده است یک شبکه DDPG انجام خواهد شد.

در ساختار دوم با عامل های جداگانه، هر ایجنت به‌طور مستقل آموزش داده می‌شود و فقط داده‌های مرتبط با زیرسیستم خود را دریافت می‌کند. در این حالت در کانال طولی از یک شبکه DDPG استفاده می‌شود که ورودی‌های آن فاصله و سرعت خودروی جلویی و همچنین محاسبه سرعت و موقعیت خودروی اصلی توسط سنسورهای خودرو است و خروجی آن شتاب طولی مناسب جهت جلوگیری از برخورد است. همچنین در کانال عرضی از یک شبکه DQN استفاده می‌شود که ورودی‌های آن خطای فاصله عرضی از مسیر مرجع و خطای سمت خودرو است، که به وسیله آن خروجی زاویه فرمان مناسب را برای خودرو تولید می‌کند. پس از آموزش هر دو ساختار، ایجنت‌ها در شرایط یکسان آزمایش شدند و نتایج بر اساس معیارهای اصلی و سرعت همگرایی و محاسبه خطاها از مقادیر مرجع ارائه می‌شوند.

۶. راه‌حل مسئله

در این روش، خودرو با استفاده از تعامل با محیط، بازخورد دریافت می‌کند و با تجربه و آموزش، یک راهبرد کنترل بهینه را یاد می‌گیرد.

معادلات ریاضی مرتبط با یادگیری تقویتی در کنترل خودرو بدون سرنشین به‌صورت زیر است:

- وضعیت محیط: وضعیت فعلی خودرو. می‌تواند یک بردار شامل اطلاعاتی مانند موقعیت، سرعت و جهت خودرو باشد.
- عمل: عملی که توسط خودرو انجام می‌شود. مثلاً فرمان دادن به خودرو برای چرخش به سمت چپ یا راست، تغییر سرعت و غیره.
- سیاست: سیاست کنترلی که توسط خودرو برای انتخاب عمل در هر وضعیت استفاده می‌شود. سیاست می‌تواند یک تابع یا قاعده‌ای باشد که وضعیت را به عمل تبدیل می‌کند.
- پاداش: پاداشی که خودرو بر اساس عمل انجام‌شده در وضعیت فعلی دریافت می‌کند. پاداش می‌تواند بر اساس معیارهایی مانند سرعت، صفحه‌بندی خط‌ها، ایمنی و غیره تعریف شود.
- تابع ارزش: تابع ارزش عمل که نشان می‌دهد چقدر عمل aدر وضعیت s مورد ارزش قرار می‌گیرد. تابع ارزش می‌تواند

به‌صورت جدولی (Q-table) یا تابعی (Q-function) پیاده‌سازی شود.

معادله اصلی یادگیری تقویتی، معادله بلمن است که به‌صورت زیر است:

$$Q^*(s,a) = Q(s,a) + \alpha * [r + \gamma * \max(Q(s',a')) - Q(s,a)] \quad (۲۵)$$

در اینجا α نرخ یادگیری است که نشان می‌دهد پاداش در به‌روزرسانی معادله بلمن دارد، چقدر تأثیر دارد. γ ضریب تخفیف است که نشان می‌دهد چقدر عوامل آینده در محاسبه تابع ارزش عمل حالت فعلی مؤثر هستند.

از این معادله برای به‌روزرسانی تابع ارزش عمل در هر مرحله استفاده می‌شود. با تجمیع تجربه‌های بیشتر، تابع ارزش عمل بهبود می‌یابد و خودرو به‌تدریج یک سیاست کنترل بهینه را یاد می‌گیرد.

با استفاده از این معادله، خودرو بدون سرنشین می‌تواند با تعامل با محیط و تجربه، یک راهبرد کنترل بهینه را برای رانندگی خود به دست آورد.

زمانی که از یک عامل یادگیری تقویتی برای چند هدف استفاده گردد، به این معنی است که ما به یک یادگیرنده مطالب متنوعی را بیاموزیم. این عمل ممکن است باعث شود که عامل بتواند در موضوعات مختلفی تصمیم بگیرد، اما باید به این نکته توجه کرد که اگر به یک عامل یک موضوع خاص آموزش داده شود و سپس همان موضوع از عامل خواسته شود، پاسخ بهتری دریافت می‌شود. چراکه با طراحی یک عامل ما داریم از تمام ظرفیت موجود در آن برای یک هدف مشخص استفاده می‌کنیم. زمانی که بخواهیم از یک منابع محدود برای چندین مقصود استفاده کنیم قطعاً کیفیت پاسخ کاهش می‌یابد؛ بنابراین درصورتی‌که برای دو هدف کنترل شتاب طولی و همچنین شتاب جانبی از دو عامل متفاوت برای هر هدف مجزا استفاده نماییم انتظار داریم که پاسخ به‌صورت دقیق‌تری داده شود.

با جدا کردن بخش کنترل شتاب طولی و کنترل شتاب جانبی در دو عامل مختلف می‌توان تا حد زیادی به سرعت الگوریتم کمک کرد. طبیعی است زمانی که عامل قصد دارد دو موضوع مختلف را مدیریت کند به شبکه‌ی عمیق‌تر و تعداد نورون‌های بیشتری احتیاج هست؛ اما زمانی که عامل یک مسئولیت خاص را بر عهده داشته باشد، آموزش آن ساده‌تر و سریع‌تر خواهد بود. همچنین با جدا شدن عامل‌ها تخمین تابع پاداش برای هر بخش کار ساده‌تری است تا زمانی که یک تابع پاداش بخواهد چندین عملکرد آن را تحت تأثیر خود قرار دهد.

۷. شبیه‌سازی و اجرا

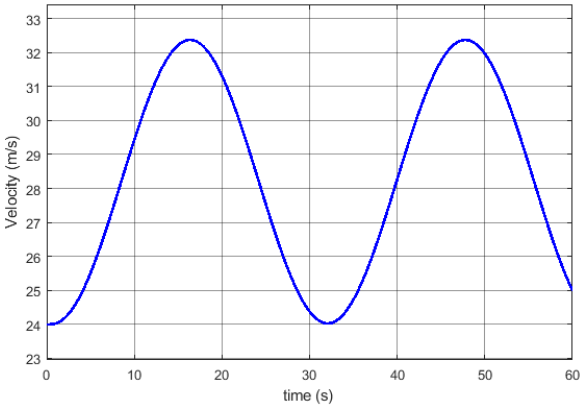
در این بخش دو مثال ارائه می‌شود و تفاوت های استفاده از عامل یکسان و عامل های جدا در بحث خودروی بدون سرنشین بررسی می‌گردد.

مثال ۱- جهت آزمایش یادگیری شبکه عصبی برای کنترل شتاب جانبی به جهت حفظ مسیر، همچنین کنترل سرعت طولی خودرو برای جلوگیری از برخورد با خودروی جلویی، از یک بلوک یادگیری تقویتی در شبیه‌ساز متلب استفاده می‌کنیم. اضافه کردن کنترل سرعت طولی خودرو باعث افزایش کارایی و ایمنی خودروی بدون سرنشین می‌گردد و برای این بهبود باید هزینه افزایش پیچیدگی محاسبات را بپذیریم. مقادیر اولیه برای شبیه-سازی سناریو اول به‌صورت جدول ۲ در نظر گرفته می‌شود.

جدول ۲. شرایط اولیه جهت شبیه سازی مثال ۱ و ۲

نام پارامتر	واحد	مقدار
C_r	---	۳۳۰۰۰
C_f	---	۱۹۰۰۰
L_r	m	۱,۶
L_f	m	۱,۲
I_z	$kg.m^2$	۲۸۷۵
V_lead	$m/\sec$	۲۴
V_ego	$m/\sec$	۱۷
m	kg	۱۵۷۵

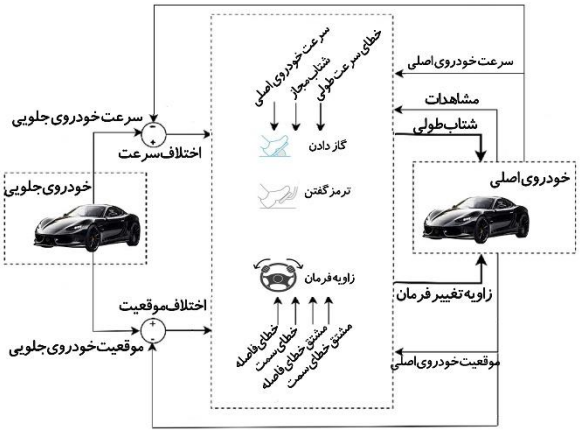
نمودار تغییرات سرعت خودروی پیش‌رونده، در شکل ۵ نشان داده شده است.



شکل ۵. نمودار تغییرات سرعت خودروی جلویی

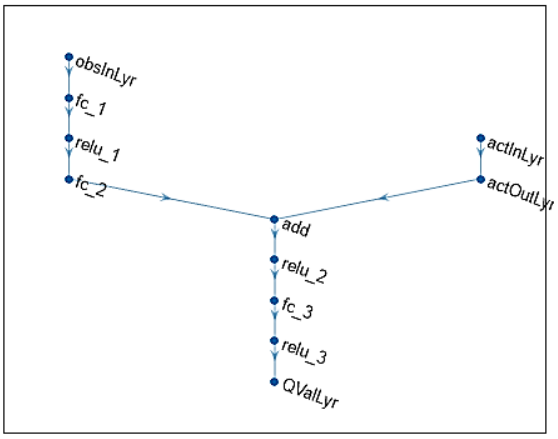
مسیر مرجع جهت خودرو در این سناریو، یک دایره واحد با شعاع ۱۰۰۰ متر ($rho = 1/R = 0.001$) خواهد بود. درنتیجه با داشتن ساختار حلقه کنترلی به‌صورت شکل ۶ و واردکردن مدل خودروی بدون سرنشین هدف[۳۰] (**EV**) و نیز مدل حرکت خودروی جلویی[۳۱] (**LV**) و تابع پاداش به‌صورت ضریبی از پارامترهای سیستم در نظر گرفته می‌شود.

باید توجه داشت که همانند خودروهای واقعی، در شبیه‌سازی باید محدودیتی برای شتاب خودرو و همچنین زاویه چرخش لاستیک جلو در نظر گرفت. این عمل به واقعی‌تر شدن شبیه‌سازی کمک زیادی می‌کند. به همین دلیل در این مثال فرض می‌شود شتاب خودرو در بازه $-3 < a < 2$ و محدودیت زاویه چرخش لاستیک‌ها توسط فرمان در محدوده $-0.5 < \delta < 0.5$ رادیان قرار داشته باشد.



شکل ۶. ساختار یادگیری تقویتی عمیق خودروی بدون سرنشین با در نظر گرفتن یک عامل مشترک

ساختار شبکه عصبی در نظر گرفته برای عامل از نوع DDPG به‌صورت شکل ۷ در نظر گرفته شده است:



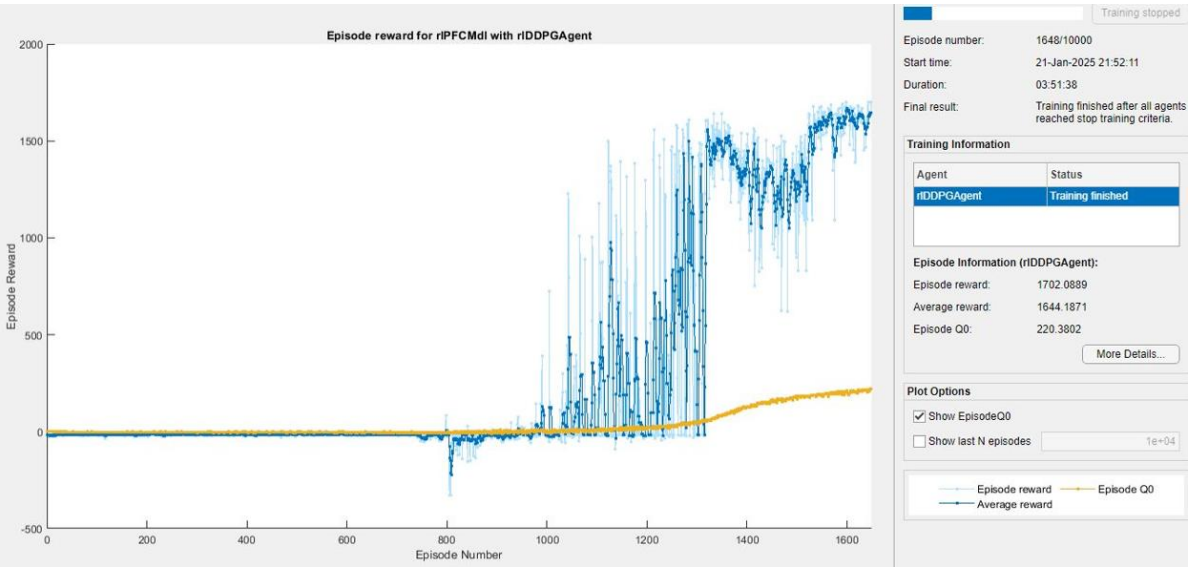
شکل ۷. شبکه عصبی مورد استفاده در DDPG

سرعت مرجع (V_{ref}) در این مثال به‌صورت زیر تعریف شده است. اگر فاصله نسبی کمتر از فاصله ایمن (D_{def}) باشد، خودرو کمترین سرعت تنظیم شده را ردیابی می‌کند. به‌این ترتیب، خودرو فاصله‌ی خود را با خودروی جلویی تنظیم می‌کند. اگر فاصله نسبی بیشتر از فاصله ایمن باشد، خودرو سرعت تنظیم شده توسط راننده (V_{ego}) را ردیابی می‌کند. در این مثال، فاصله ایمن به‌عنوان یک تابع خطی از سرعت طولی خودرو به‌صورت $t_{gap} * V + D_{def}$ تعریف شده است که $D_{def} = 40$ است.

تابع پاداش در نظر گرفته شده برای این عامل از رابطه زیر به دست می‌آید:

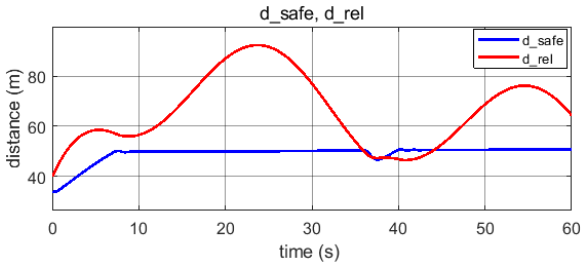
$$R = -(100e_1^2 + 500u_{t-1}^2 + 10e_v^2 + 100a_{t-1}^2) \times e_2^{-3} - 10F_t + 2H_t + M_t \quad (۲۶)$$

که در آن $e_v = V_{ref} - V_{ego}$ و پارامترهای شرطی F_t، H_t و M_t به‌صورت زیر مقدار می‌گیرند. اگر آموزش شبکه به محدودیت‌های در نظر گرفته شده برای خطاها روبرو شد و متوقف شد، مقدار F_t برابر ۱ در غیر این صورت صفر است. درصورتی‌که خطای فاصله کوچک‌تر از ۰٫۰۱ شود مقدار H_t برابر ۱ و در غیر این صورت صفر است. درصورتی‌که $e_v{}^2 < 1$ باشد M_t برابر ۱ و در غیر این صورت مقدار آن صفر است. با شروع فرایند آموزش توسط جعبه‌ابزار آموزش عامل یادگیری تقویتی در نرم‌افزار متلب، نتایج به‌صورت شکل ۸ به دست می‌آید:



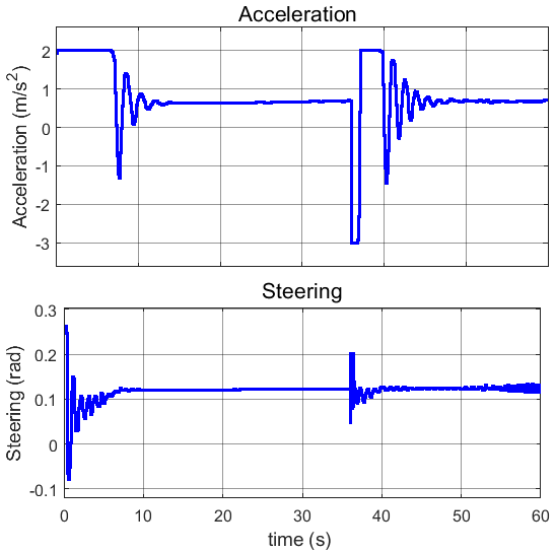
شکل ۸. نحوه پیشرفت یادگیری شبکه با افزایش مقدار پاداش

مشاهده می‌شود که برای آموزش شبکه عصبی عمیق با یک عامل به میزان حدود ۴ ساعت موردنیاز است و تا تکرارهای ۱۰۰۰ شبکه رشدی را از خود نشان نمی‌دهد. سپس تا تکرار ۱۳۶۰ شبکه سناریوهایی را تجربه می‌کند که در آن به پاداش بالاتری دست پیدا کرده است ولی پایدار نیست. این رفتار ناشی از آن است که شبکه هنوز متوجه نشده است که بالا رفتن پاداش مربوط به تغییر کدام وزن یا هایپر پارامتر است. درنهایت با تکرارهای بیشتر به پایداری در پاداش بالای شبکه می‌رسد. ناگفته نماند که اگر ساختار شبکه با نوع مسئله متناسب نباشد، یا مقدار ایدئال پاداش نامناسب انتخاب شود، شبکه به جواب نمی‌رسد و روند نمودار آموزش آن به طرفی پیش می‌رود که مقدار پاداش‌ها منفی‌تر شده و بهبودی در رفتار سیستم رخ نمی‌دهد. کنترل سرعت طولی خودرو، طوری تنظیم شده است تا فاصله‌ی ایمن را متناسب با خودروی جلویی تنظیم کند و مقدار فاصله ایمن را با میزان فاصله از خودروی جلویی تطبیق می‌دهد. شکل ۹ مقدار فاصله دو خودرو از یکدیگر (خط قرمز) و همچنین میزان فاصله ایمن بین این دو خودرو (خط آبی) را در طول زمان شبیه‌سازی ۶۰ ثانیه نشان می‌دهد.



شکل ۹. میزان تغییرات فاصله نسبی دو خودرو به همراه فاصله ایمن متناسب با سرعت خودرو

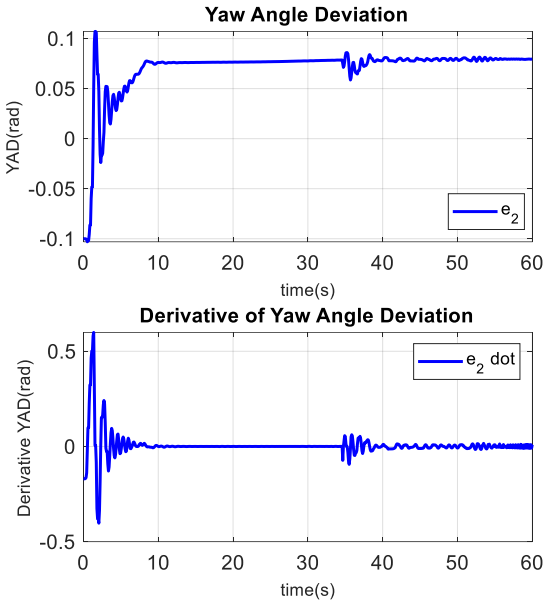
در این شکل مشاهده می‌شود زمانی که شتاب خودروی جلویی منفی است و فاصله دو خودرو از یکدیگر محدوده مجاز را رد می‌کند، به علت کاهش ناگهانی خودروی هدف، ابتدا فاصله ایمن متناسب با سرعت خودرو به‌طور ناگهانی کاهش و سپس با پایدارتر شدن تغییرات فاصله دوباره به مقدار فاصله تعیین‌شده برای آن بازمی‌گردد. مقدار شتاب طولی و مقدار زاویه فرمان جهت رسیدن به هدف، به‌صورت شکل ۱۰ نشان داده شده است.



شکل ۱۰. مقدار تغییرات شتاب و زاویه فرمان خودرو طی مسیر

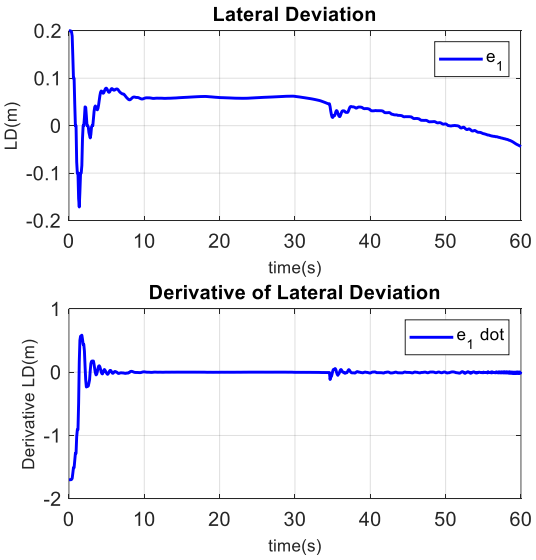
در اواسط نمودار در حدود زمان ۳۳ ثانیه پس از گذشت شروع شبیه‌سازی، تغییر ناگهانی در مقدار شتاب و زاویه فرمان رخ می‌دهد. این تغییر ناشی از این است که فاصله خودروی جلویی از خودروی هدف به میزانی کمتر از حد استاندارد رسیده است. پس برای عدم برخورد خودروها، احتیاج به کاهش شتاب از طریق ترمز شده است؛ که تغییر ناگهانی سرعت روی زاویه فرمان نیز تأثیر گذاشته است. تأثیر شتاب طولی بروی تغییرات زاویه فرمان ناشی از کنترل این دو پارامتر با یک عامل یادگیری تقویتی است. تغییرات ناگهانی وجود آمده در شتاب، باعث تغییر در تابع پاداش شده و عامل یادگیری تقویتی برای بالا بردن دوباره آن دست به تغییر پارامترهای تحت کنترل خود میزند. درنتیجه مشاهده می‌شود که کنترل خودرو در این حالت در دو فاز کنترل شتاب طولی و جانبی از یکدیگر مستقل نیست.

میزان انحراف زاویه خودرو از مسیر مرجع[۳۲] (YAD) و تغییرات آن در شکل ۱۱ نشان داده شده است.



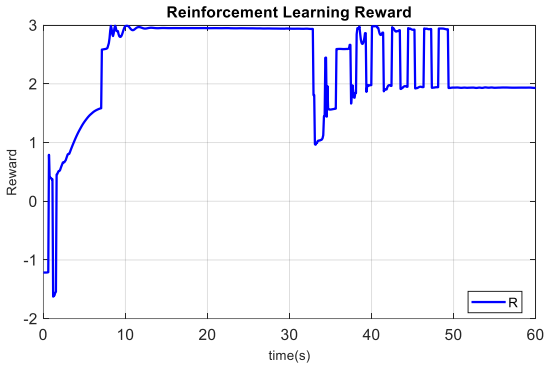
شکل ۱۱. مقدار تغییر زاویه سمت خودرو از مسیر مرجع

در این شکل مشاهده می‌شود، با توجه به مقادیر اولیه خطای در نظر گرفته شده برای خودرو، در ابتدا تا خودرو بر خطاهای اولیه غلبه کند و خود را به مسیر مشخص شده برساند، مقدار خطا از مسیر مرجع نوسانات زیادی دارد. سپس با پایدار شدن خودرو در مسیر این مقدار همچنین میزان انحراف جانبی از مسیر مرجع در شکل زیر مشاهده می‌شود.



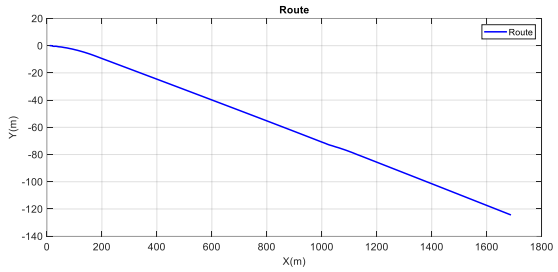
شکل ۱۲. مقدار فاصله عرضی از مسیر مرجع در طی مسیر

شکل ۱۲ فاصله اقلیدسی خودرو از مسیر مرجع (LD) [۳۳] به همراه مشتق آن نمایش می‌دهد. با توجه به وجود انحنای در مسیر خودرو، واضح است که خودرو در طی مسیر تلاش می‌کند تا فاصله خود را مسیر مرجع کاهش دهد. پس از زمان ۷ ثانیه تا زمان ۳۳ ثانیه ابتدای شبه سازی، به علت وجود شتاب طولی و محدودیت چرخش زاویه فرمان فاصله در مقدار تقریباً ثابتی قرار می‌گیرد؛ اما پس از ترمز ناگهانی خودرو و تغییر شتاب و زاویه فرمان، خودرو خود را به مسیر مرجع نزدیک می‌کند. نمودار مشتق تغییرات این خطا نشان دهنده شیب ملایم تغییرات فاصله هست. جهت بررسی بهتر مقدار تغییرات تابع پاداش در شکل ۱۳ نمایش داده شده است.



شکل ۱۳. تغییرات تابع پاداش شبکه عصبی عمیق در طی مسیر

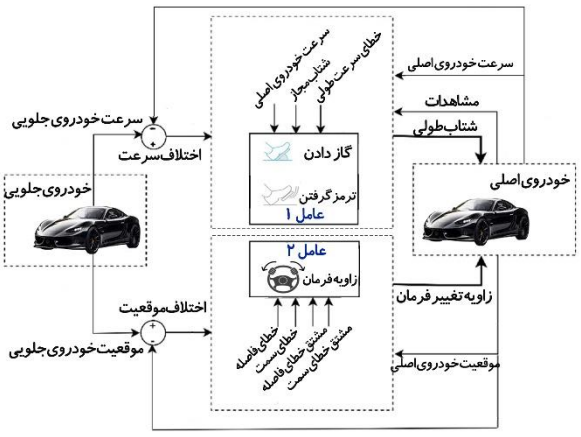
نوسانات پارامترهای کنترل شونده به نوسانات تابع پاداش مربوط می‌شود. همان‌طور که انتظار می‌رود تغییرات تابع پاداش در زمان کم شدن فاصله ایمن دو خودرو از یکدیگر شروع می‌شود و با نوسان در بازه‌ای محدود دوباره به مقدار پایدار می‌رسد. مسیر طی شده توسط خودرو را در این سناریو نیز شکل ۱۴ نشان می‌دهد.



شکل ۱۴. مسیر طی شده توسط خودروی بدون سرنشین

در مثال بعدی همین سناریو با بهبود ساختار حلقه کنترلی انجام می‌پذیرد.

مثال ۲– با جدا کردن دو هدف یعنی کنترل گاز و ترمز برای کنترل شتاب طولی و کنترل زاویه فرمان جهت کنترل شتاب جانبی خودرو در دو عامل، ساختار مسئله به‌صورت زیر درمی آید:



شکل ۱۵. ساختار یادگیری تقویتی عمیق خودروی بدون سرنشین با دو عامل مجزا

در شکل ۱۵ مشاهده می‌شود که عامل‌ها هر فاز کنترلی از یکدیگر تفکیک شده و هرکدام با شبکه‌های مخصوص به خود آموزش می‌بینند. این دو عامل درنهایت با اعمال به دینامیک خودروی

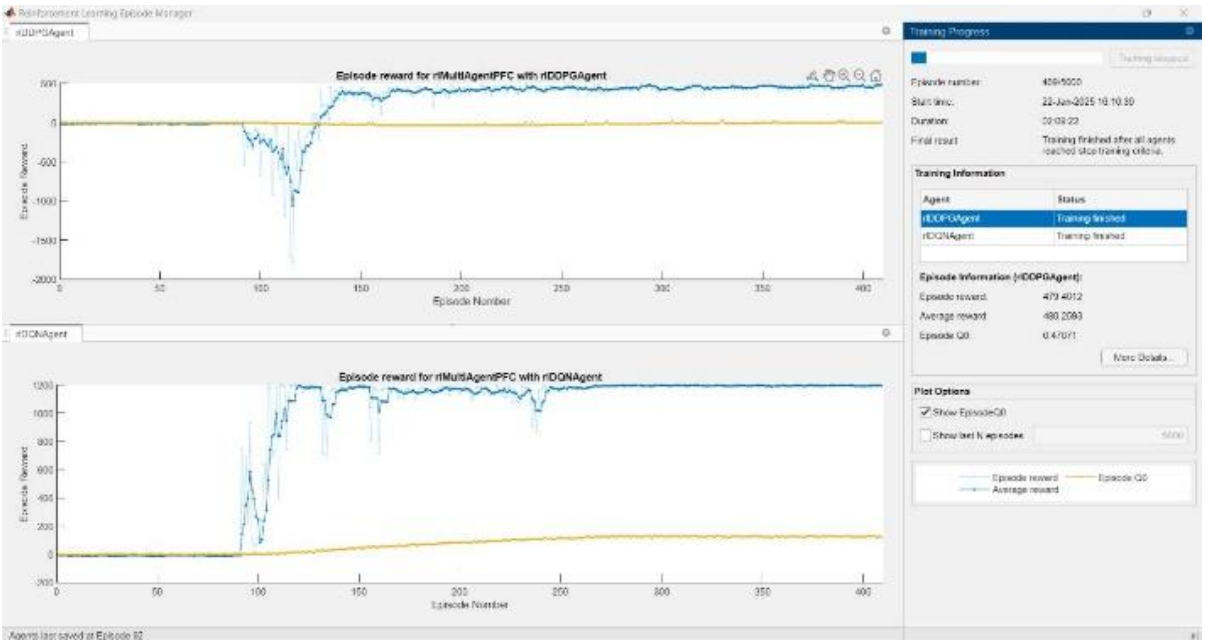
هدف با یکدیگر تبادل اطلاعات می‌کنند و به‌طور کامل از یکدیگر مستقل نیستند. شبکه در نظر گرفته شده برای فاز کنترل طولی در این مثال از نوع DQN و برای فاز کنترل شتاب جانبی از نوع DDPG هست.

تابع پاداش برای عامل‌های ۱ و ۲ به‌صورت معادلات زیر تعریف می‌شود:

$$R_1 = -(100e_1^2 + 500u_{t-1}^2) \times 0.001 - 10F_t + 2H_t \quad (۲۷)$$

$$R_2 = -(10e_v^2 + 100a_{t-1}^2) \times 0.001 - 10F_t + M_t \quad (۲۸)$$

با استفاده از جعبه‌ابزار مخصوص یادگیری تقویتی در نرم‌افزار متلب، روند آموزش شبکه‌های عصبی در دو کانال طولی و عرضی مطابق شکل ۱۶ انجام می‌گیرد.



شکل ۱۶. روند آموزش شبکه DDPG و DQN در دو کانال طولی و عرضی خودرو

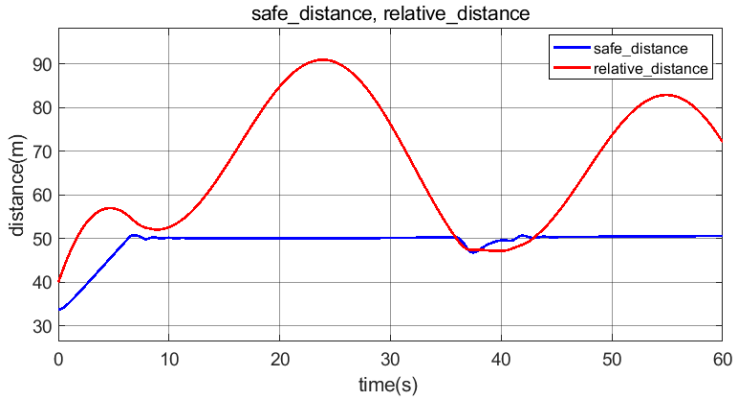
نتایج به‌دست‌آمده پس از اتمام فرآیند آموزش شبکه‌های عصبی برای دو کانال طولی و عرضی در جدول ۳ مشاهده می‌شود:

جدول ۳. پارامتر های دو شبکه عصبی مجزا آموزش دیده مثال ۲

	Reward	Q0	Stopped at (episode)	Time (h)
DQN	۱۱۹۱/۶۹	۱۲۴/۰۵۱۱	۱۱۹۴	۲/۰۹
DDPG	۴۷۹/۴۰۱	۰/۴۷۰۷۰۹	۴۸۰	۱/۴۲

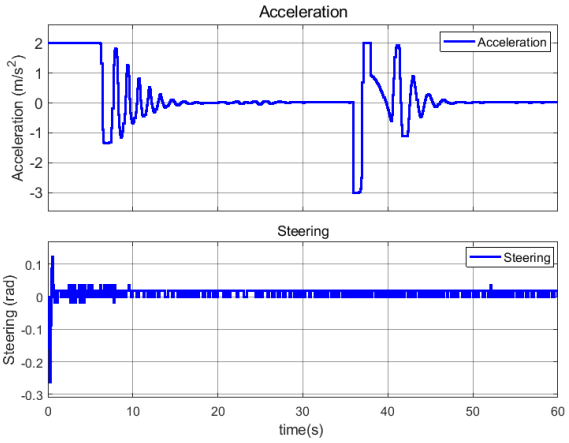
نتایج جدول بالا نشان دهنده همگرا شدن سریع‌تر در حالت عامل‌های جداشده است. شبیه‌سازی هر دو مثال با یک سیستم ثابت با پردازنده Core i5 نسل ۳ و مقدار RAM 12G در حالت بیشترین بهره‌وری کاری سیستم انجام‌شده است. روند آموزش در حالت تک‌عاملِ حدود چهار ساعت به طول انجامید، درصورتی‌که با جدا کردن عامل‌ها این زمان حدود نصف کاهش یافته است.

در شکل ۱۷ مقدار فاصله دو خودرو از یکدیگر در طول شبیه‌سازی و نیز میزان فاصله ایمن تطبیقی مشاهده می‌شود. نمودار آبی فاصله ایمن و نمودار قرمزرنگ فاصله‌ی دو خودرو را در هر لحظه از مسیر نشان می‌دهد.



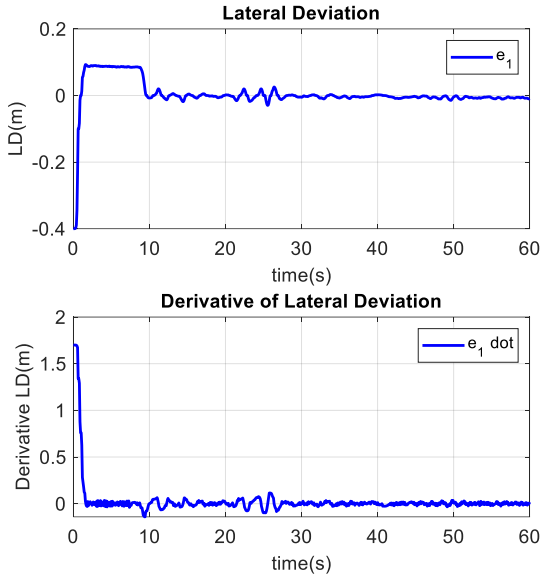
شکل ۱۷. مقدار فاصله نسبی دو خودرو از یکدیگر

همانند رویدادی که در مثال ۱ افتاد، زمانی که فاصله دو خودرو به‌طور ناگهانی یکدیگر کاهش می‌یابد، به خودرو هدف شتاب منفی القا می‌شود که به همین سبب، فاصله ایمن بین دو خودرو مطابق با سرعت خودرو ابتدا کاهش‌یافته و افزایش دوباره شتاب به مقدار تعیین‌شده بازمی‌گردد. شکل ۱۸ نیز این روند را به‌خوبی نمایش می‌دهد. نکته حائز اهمیت در اینجا تأثیر کم کانال کنترل عرضی به روی کانال کنترل طولی است. در این شکل مشاهده می‌شود که به علت جداسازی عامل‌های کنترل طولی و عرضی، زاویه چرخ‌ها تأثیر بسیار ناچیزی از تغییرات شتاب طولی گرفته است.



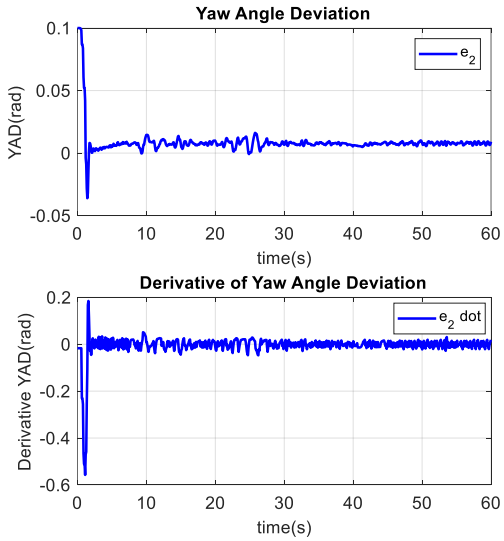
شکل ۱۸. میزان تغییرات شتاب طولی و زاویه فرمان در طی مسیر

خطای فاصله و مشتق آن نیز در شکل ۱۹ نمایش داده شده است. واضح است که خودرو در کمتر از ۱۰ ثانیه پس از شروع شبیه‌سازی در مسیر قرا می‌گیرد و با خطای ناچیزی به مسیر خود ادامه می‌دهد. قدرت کنترل این الگوریتم زمانی بهتر اثبات می‌شود که مشاهده می‌شود تغییرات ناگهانی شتاب، تأثیر چندانی روی خطای فاصله از مسیر مرجع نداشته است و خودرو توانسته به‌خوبی نوسانات حاصل از تغییرات ناگهانی شتاب را دمپ کند.



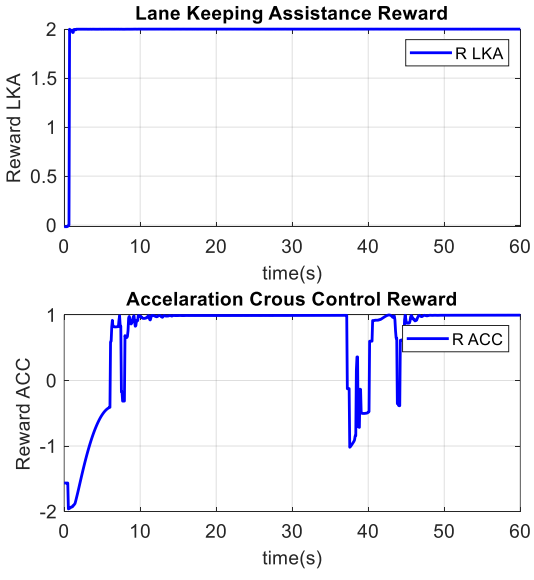
شکل ۱۹. مقدار خطای فاصله جانبی از مسیر مرجع

شکل ۲۰ نمایانگر روند تغییرات خطای سمت و مشتق آن در طول شبیه‌سازی است. با توجه به وجود انحنای ثابت در مسیر دایره مرجع، منطقی است که میزان خطای سمت به صفر نرسد. چراکه در مسیر منحنی، سمت خودرو باید کمی از زاویه سمت مسیر فاصله بگیرد تا خودرو را در مسیر نگه دارد.



شکل ۲۰. مقدار خطای زاویه سمت خودرو از مسیر مرجع

تغییرات تابع پاداش در دو کانال طولی و عرضی به ترتیب برای شبکه‌های DDPG و DQN در شکل ۲۱ نمایش داده شده است. به‌خوبی نمایان است که در کانال عرضی زودتر از کانال طولی به مقدار مطلوب رسیده است و تغییرات ناگهانی شتاب رو عملکرد آن تأثیری نداشته است.



شکل ۲۱. مقادیر تابع پاداش در دو شبکه عصبی کانال طولی و عرضی

برای بررسی بهتر دقت عملکرد الگوریتم کنترلی در ساختار بهبود یافته، در جدول ۴ مقادیر RMS[۳۴] خطا برای خطای سمت و خطای فاصله، به همراه مشتقات آن با یکدیگر مقایسه شده است. نتایج به‌دست‌آمده عملکرد بهتر ساختار کنترلی دوم را اثبات می‌کند.

جدول ۴. مقایسه مقدار RMS برای خطای سمت و خطای فاصله و مشتق خطاها

RMS	YAD(rad)	LD(m)	D_YAD(rad)	D_LD(m)
One Agent	۰/۰۷۵۴	۰/۰۵۱۶	۰/۰۶۶۵	۰/۲۲۲۹
Two Agent	۰/۰۱۴۶	۰/۰۴۰۶	۰/۰۴۱۱	۰/۲۰۰۴

در جدول ۳، با توجه به مقدار میانگین خطاها، می‌توان عملکرد بهتر و دقیق‌تر روش استفاده شده در مثال دوم را نتیجه گرفت.

۸. نتیجه‌گیری و پیشنهادها

این پژوهش به بررسی تأثیر تفکیک عامل‌های یادگیری تقویتی بر کنترل شتاب طولی و جانبی خودروهای بدون سرنشین پرداخته است. نتایج شبیه‌سازی‌ها نشان می‌دهند که پس از جدا شدن عامل‌ها، خطاهای مربوط به جهت و فاصله به‌طور قابل توجهی کاهش یافته است. این بهبود در دقت عملکرد در حالی حاصل‌شده است که زمان آموزش شبکه به‌مراتب کمتر از حالت استفاده از یک عامل مشترک بوده است. این یافته‌ها نشان‌دهنده کارایی بالای رویکرد پیشنهادی در مدیریت پیچیدگی‌های کنترل خودروهای خودران هستند و اهمیت تفکیک وظایف را در بهبود عملکرد سیستم‌های هوشمند تأکید می‌کنند .با توجه به نتایج مثبت این تحقیق، می‌توان نتیجه گرفت که طراحی سیستم‌های کنترل با استفاده از عامل‌های مستقل نه‌تنها منجر به افزایش دقت و سرعت آموزش می‌شود، بلکه قابلیت انطباق با شرایط مختلف جاده‌ای را نیز بهبود می‌بخشد. این رویکرد می‌تواند به‌عنوان یک چارچوب مؤثر برای توسعه فناوری‌های پیشرفته در صنعت خودروهای خودران با عملکرد بالا مورد استفاده قرار گیرد.

در ادامه پیشنهاد می‌شود که این الگوریتم با الگوریتم‌های کنترل مقاوم ترکیب شود. این موضوع باعث می‌شود که وجود نامعینی‌ها در پارامترهای خودروی بدون سرنشین تأثیر مخربی

روی کنترل طولی و عرضی نداشته باشد و به پایداری خودرو کمک کند. همچنین تحقیق بر روی الگوریتم‌های یادگیری تقویتی پیشرفته‌تر و تکنیک‌های بهینه‌سازی برای افزایش سرعت همگرایی و دقت کنترل نیز می‌تواند زمینه‌ساز نتایج بهتر باشد.

۹. پی‌نوشت‌ها

1. World Health Organization (WHO)
2. Vulnerable Road User (VRU)
3. Focus Group on AI for Autonomous and Assisted Driving (FG-AI4AD)
4. Autonomous vehicle (AV)
5. Vehicle to Everything (V2X)
6. Internet of Things (IoT)
7. Intelligent Transport System (ITS)
8. natural language processing (nlp)
9. Reinforcement Learninig (RL)
10. motion planning
11. Artifcial Intelligence (AI)
12. Human- Artifcial Intelligence (H-AI)
13. Differential of Variable Speed Limits (DVSL)
14. *Deep* Q-Network (DQN)
15. Proximal Policy Optimization (PPO)
16. *Deep* Deterministic Policy Gradient (DDPG)
17. *Inverse* Reinforcement Learninig (IRL)
18. Agent
19. Degree of Freedom
20. Dynamic programming (DP)
21. Time-course *Deep Learning (TDL)*
22. Markov decision process
23. State
24. Action
25. Reward
26. Policy
27. Value
28. Partially observable markov decision process (POMDP)
29. Shallow machine learning
30. Ego Vehicle (EV)
31. Lead Vehicle (LV)
32. Yaw Angel Deviation (YAD)
33. Lateral Deviation (LD)
34. Root Mean Square (RMS)

۱۰. مراجع

- Artuñedo, A., Moreno-Gonzalez, M., & Villagra, J. (2024). “Lateral control for autonomous vehicles: A comparative evaluation”. Annual reviews in control, 57, 100910.

- Baheri, A., Kolmanovsky, I., Girard, A., Tseng, H. E., & Filev, D. (2020, July) “Vision-based autonomous driving: A model learning approach”. In 2020 American Control Conference (ACC) (pp. 2520-2525). IEEE.

- Bay, O. (2021) “Abi research forecasts 8 million vehicles to ship with sae level 3, 4 and 5 autonomous technology in 2025”.

- Bellotti, F., Lazzaroni, L., Capello, A., Cossu, M., De Gloria, A., & Berta, R. (2023) “Explaining a deep reinforcement learning (DRL)-based automated driving agent in highway simulations”. IEEE Access, 11, 28522-28550.

- Cao, Y., Ni, K., Jiang, X., Kuroiwa, T., Zhang, H., Kawaguchi, T., & Jiang, W. (2023). Path following for autonomous ground vehicle using DDPG algorithm: a reinforcement learning approach. Applied Sciences, 13(11), 6847.

- Chen, J., Zhang, C., Luo, J., Xie, J., & Wan, Y. (2020). “Driving maneuvers prediction based autonomous driving control by deep Monte Carlo tree search”. IEEE transactions on vehicular technology, 69(7), 7146-7158.

- Claussmann, L., Revilloud, M., Gruyer, D., & Glaser, S. (2019) “A review of motion planning for highway autonomous driving”. IEEE Transactions on Intelligent Transportation Systems, 21(5), 1826-1848.

- Du, Y., Chen, J., Zhao, C., Liao, F., & Zhu, M. (2023). “A hierarchical framework for

-Hoveidafard, A., Fardmoradinia, S., Golchin, B. and Ghaffari, A. (2023) "Simulation of autonomous vehicles using machine learning methods". International Journal of Transportation Engineering, 15(2), 3483-3507 (in persian).

- Huang, Z., Zhang, J., Tian, R., & Zhang, Y. (2019) "End-to-end autonomous driving decision based on deep reinforcement learning". In 2019 5th International Conference on Control, Automation and Robotics (ICCAR) (pp. 658-662). IEEE.

-Kuutti, S., Bowden, R., Jin, Y., Barber, P., & Fallah, S. (2020) "A survey of deep learning applications to autonomous vehicle control". IEEE Transactions on Intelligent Transportation Systems, 22(2), 712-733.

- Lamssaggad, A., Benamar, N., Hafid, A. S., & Msahli, M. (2021) "A survey on the current security landscape of intelligent transportation systems". IEEE Access, 9, 9180-9208.

- Lee, K., Isele, D., Theodorou, E. A., & Bae, S. (2022) "Spatiotemporal costmap inference for MPC via deep inverse reinforcement learning". IEEE Robotics and Automation Letters, 7(2), 3194-3201.

- Li, Z., Yuan, S., Yin, X., Li, X., & Tang, S. (2023) "Research into autonomous vehicles following and obstacle avoidance based on deep reinforcement learning method under map constraints". Sensors, 23(2), 844.

-Liu, Y., & Diao, S. (2024) "An automatic driving trajectory planning approach in complex traffic scenarios based on integrated driver style inference and deep reinforcement learning". PLoS one, 19(1), e0297192.

- Ma, Y., Wang, Z., Yang, H., & Yang, L. (2020) "Artificial intelligence applications in improving ride comfort of autonomous vehicles via deep reinforcement learning with external knowledge". Computer-Aided Civil and Infrastructure Engineering, 38(8), 1059-1078.

- Ebrahimi, M. (2023). "Broken rail detection with texture image processing using two-dimensional gray level co-occurrence matrix". arXiv preprint arXiv:2304.11592.

- Ebrahimi, M., & Asgari, M. (2021). "Robust fractional-order fixed-structure controller design for uncertain non-commensurate fractional plants using fractional Kharitonov theorem". Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 235(8), 1375-1387.

- Ebrahimi, M., & Nasrollahi, S. (2025). "Fractional Guidance Law with Impact Angle Constraint and Seeker's Look Angle Limits". Unmanned Systems, 13(01), 261-277.

- El Hamdani, S., Benamar, N., & Younis, M. (2020) "Pedestrian support in intelligent transportation systems: challenges, solutions and open issues". Transportation research part C: emerging technologies, 121, 102856.

- Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., ... & Dietmayer, K. (2020) "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges". IEEE Transactions on Intelligent Transportation Systems, 22(3), 1341-1360.

- Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020) "A survey of deep learning techniques for autonomous driving". Journal of field robotics, 37(3), 362-386.

- Guo, J., Cheng, S., & Liu, Y. (2020) "Merging and diverging impact on mixed traffic of regular and autonomous vehicles". IEEE Transactions on Intelligent Transportation Systems, 22(3), 1639-1649.

the development of autonomous vehicles: A survey". IEEE/CAA Journal of Automatica Sinica, 7(2), 315-329.

- Mahaadevan, V. C., Narayanamoorthi, R., Panda, S., Dutta, S., & Dooly, G. (2024) "AViTRoN: Advanced vision track routing and navigation for autonomous charging of electric vehicles". IEEE Access.

- Mao, Z., Liu, Y., & Qu, X. (2024) "Integrating big data analytics in autonomous driving: An unsupervised hierarchical reinforcement learning approach". Transportation Research Part C: Emerging Technologies, 162, 104606.

- Ning, H., Yin, R., Ullah, A., & Shi, F. (2021) "A survey on hybrid human-artificial intelligence for autonomous driving". IEEE Transactions on Intelligent Transportation Systems, 23(7), 6011-6026.

- Peličić, D., Ristić, B., & Radević, S. (2024) "Factors for the occurrence of road traffic injuries and better care of injured persons". Zdravstvena zaštita, 53, 59.

- Pérez-Gil, Ó., Barea, R., López-Guillén, E., Bergasa, L. M., Gómez-Huélamo, C., Gutiérrez, R., & Díaz-Díaz, A. (2022) "Deep reinforcement learning based control for Autonomous Vehicles in CARLA". Multimedia Tools and Applications, 81(3), 3553-3576.

- Rajamani, R. (2006). Vehicle dynamics and control. Boston, MA: Springer US.

- Singh, S. (2015) "Critical reasons for crashes investigated in the national motor vehicle crash causation survey" (No. DOT HS 812 115).

- Toromanoff, M., Wirbel, E., & Moutarde, F. (2020) "End-to-end model-free reinforcement learning for urban driving using implicit affordances". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7153-7162).

- Van Hasselt, H., Guez, A., & Silver, D. (2016) "Deep reinforcement learning with double q-learning". In Proceedings of the AAAI conference on artificial intelligence (Vol. 30, No. 1).

- Wu, Y., Tan, H., Qin, L., & Ran, B. (2020) "Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm". Transportation research part C: emerging technologies, 117, 102649.

- Wang, C., Cui, X., Zhao, S., Zhou, X., Song, Y., Wang, Y., & Guo, K. (2024) "A deep reinforcement learning-based active suspension control algorithm considering deterministic experience tracing for autonomous vehicle". Applied Soft Computing, 153, 111259.

- Ye, F., Cheng, X., Wang, P., Chan, C. Y., & Zhang, J. (2020) "Automated lane change strategy using proximal policy optimization-based deep reinforcement learning". In 2020 IEEE Intelligent Vehicles Symposium (IV) (pp. 1746-1752). IEEE.

- Yuan, M., Shan, J., & Mi, K. (2023) "From Naturalistic Traffic Data to Learning-Based Driving Policy: A Sim-to-Real Study". IEEE Transactions on Vehicular Technology.

- Yusuf, S. A., Khan, A., & Souissi, R. (2024) "Vehicle-to-everything (V2X) in the autonomous vehicles domain–A technical review of communication, sensor, and AI technologies for road user safety". Transportation Research Interdisciplinary Perspectives, 23, 100980.

- Zhang, Y., Sun, P., Yin, Y., Lin, L., & Wang, X. (2018) "Human-like autonomous vehicle speed control by deep reinforcement learning

with double Q-learning". In 2018 IEEE intelligent vehicles symposium (IV) (pp. 1251-1256). IEEE.

- Zhu, M., Wang, Y., Pu, Z., Hu, J., Wang, X., & Ke, R. (2020) "Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving". Transportation Research Part C: Emerging Technologies, 117, 102662.

محسن ابراهیمی مدرک کارشناسی ارشد مهندسی برق را در سال ۱۳۹۴ از دانشگاه صنعتی شاهرود، شاهرود، ایران دریافت کرد. او در حال حاضر مشغول تحصیل در مقطع دکترا در دانشکده مهندسی برق از دانشگاه صنعتی مالک اشتر، تهران، ایران است. زمینه‌های تحقیقاتی مورد علاقه او شامل یادگیری تقویتی، کنترل تطبیقی، کنترل مقاوم، حساب دیفرانسیل و انتگرال کسری و الگوریتم‌های یادگیری ماشین است. او پروژه‌هایی در زمینه اینترنت اشیا و خانه‌های هوشمند انجام داده است.

فیروز اللهوردی زاده، متولد پنجم خرداد ۱۳۴۷، ایران اردبیل، فارغ التحصیل دبیرستان هدف شماره ۱ تهران در رشته ریاضی فیزیک در سال ۱۳۶۵، لیسانس مهندسی برق مخابرات از دانشگاه علم و صنعت ایران، خرداد ۱۳۷۰، فارغ التحصیل مهندسی برق کنترل از دانشگاه صنعتی امیر کبیر، شهریور ۱۳۷۴، فارغ التحصیل دکتری تخصصی هوافضا دینامیک پرواز وکنترل از دانشگاه خواجه نصیرالدین طوسی، اردیبهشت ۱۳۹۷، استادیار دانشگاه صنعتی مالک اشتر، تهران، ایران است. علاقه مند به تحقیق در حوزه هوش مصنوعی و کاربردهای آن، متخصص هدایت و کنترل ناوبری، آزمایشگاه HWIL، طراحی سیستمهای پیچیده، با بیش از ۳۰ سال سابقه علمی و تخصصی، تهیه وتدوین دهها گزارش فنی وتخصصی، بیش از ده مقاله علمی پژوهشی و ISI.

عبدالرضا کاشانی نیا مدرک کارشناسی خود را در رشته مهندسی برق از دانشگاه تهران در سال ۲۰۰۱ و مدارک کارشناسی ارشد و دکترای خود را به ترتیب در سال‌های ۲۰۰۳ و ۲۰۱۱ از دانشگاه صنعتی امیرکبیر، تهران، ایران دریافت کرد. از سال ۲۰۱۲، او استادیار دانشکده مهندسی برق، دانشگاه صنعتی مالک اشتر، تهران، ایران است. علایق تحقیقاتی او شامل کنترل غیرخطی و تطبیقی، کنترل بهینه و پیش‌بین، سیستم‌های چندعامله و کنترل تحمل‌پذیر خطا است.