

الگوریتم ناوبری رفتار مبنا برای دوچرخه خودپایدارساز در حضور موانع به کمک یادگیری تقویتی عمیق

آرش کریمی^۱، مجید ساده‌دل^{۲*}

۱- گروه مهندسی مکاترونیک، دانشگاه تربیت مدرس، تهران، ایران

۲- گروه مهندسی مکانیک، دانشگاه تربیت مدرس، تهران، ایران

پست الکترونیکی نویسندگان:

۱- arash_karimi@modares.ac.ir

۲- majid.sadedel@modares.ac.ir

چکیده:

خودروهای دوچرخه به دلیل ابعاد کوچک و مانورپذیری بالا در فضاها محدود کاربرد گسترده‌ای دارند؛ باین‌حال، حفظ تعادل آن‌ها در سرعت‌های کمتر از پنج متر بر ثانیه چالشی اساسی است. در این پژوهش، مدل‌سازی دینامیکی دوچرخه‌ی خودپایدارساز مجهز به چرخ عکس‌العمل با رویکرد لاگرانژی در متلب انجام و با مدل چندجسمی سیم‌اسکیپ اعتبارسنجی شد. جهت بهبود پایداری در سرعت‌های پایین، دو الگوریتم یادگیری تقویتی عمیق شامل روش‌های گرادینان خط‌مشی قطعی عمیق و نوع دوگانه‌ی تأخیری آن برای بهینه‌سازی گشتاور ترمیمی به کار رفتند. الگوریتم ناوبری رفتارمبنا نیز توسط کنترلگر تناسبی-انترالی-مشتقی با حلقه تعادل ادغام شد. ارزیابی مقایسه‌ای با تنظیم‌کننده خطی مربعی در سه سناریوی مسیر مستقیم، منحنی با مانع و اغتشاش جاده نشان داد روش دوگانه‌ی تأخیری با غلبه بر دینامیک غیرخطی، نوسانات کمتری دارد. این رویکرد در ناوبری بین موانع، ریشه‌ی میانگین مربعات خطای غلتش را به ۱.۲۳ درجه کاهش داد که راه‌حلی کارآمد برای خودروهای خودران است.

واژگان کلیدی:

کنترل تعادل، دوچرخه خودپایدارساز، الگوریتم ناوبری رفتار مبنا، چرخ عکس‌العمل، یادگیری تقویتی.

* مجید ساده‌دل، دانشیار دانشکده مهندسی مکانیک، دانشگاه تربیت مدرس.

ایمیل: majid.sadedel@modares.ac.ir (نویسنده مسئول مقاله)

Behavior-Based Navigation Algorithm of a Self-Balancing Bicycle in the Presence of Obstacles Using Deep Reinforcement Learning

Arash Karimi¹, Majid Sadedel²

¹ Mechatronic Engineering Department, Tarbiat Modares University, Tehran, Iran

² Mechanical Engineering Department, Tarbiat Modares University, Tehran, Iran

Abstract:

Two-wheeled vehicles, such as autonomous bicycles and motorcycles, are widely utilized due to their compact dimensions, low inertia, and high maneuverability. These characteristics make them highly suitable for efficient navigation in confined spaces. However, maintaining their dynamic balance, particularly at speeds below five meters per second, remains a significant challenge. At such low velocities, the natural stabilizing effects of gyroscopic precession and caster geometry diminish, rendering the vehicle highly unstable. To address this instability, this study presents the dynamic modeling of an autonomous self-balancing bicycle equipped with a reaction wheel. The nonlinear equations of motion were developed based on the Lagrangian formulation within the MATLAB/Simulink environment. The derived state-space representation was thoroughly validated against a Simscape multibody simulation under identical pulse torque excitation, confirming the accuracy of the mathematical framework. To enhance dynamic stability at low speeds, two deep reinforcement learning (DRL) algorithms were implemented: the Deep Deterministic Policy Gradient (DDPG) and the Twin Delayed Deep Deterministic Policy Gradient (TD3). These intelligent, model-free controllers were designed to continuously optimize the reaction wheel's restorative torque commands. Furthermore, a behavior-based navigation algorithm, capable of goal-reaching and obstacle-avoidance, was integrated with the primary balance loop utilizing a Proportional-Integral-Derivative (PID) steering controller. An extensive comparative evaluation was conducted against a classical Linear Quadratic Regulator (LQR) across three operational scenarios: straight-path balance, curved-path navigation involving obstacles, and disturbance rejection over a speed-bump. The simulation results demonstrated that the TD3 algorithm outperformed conventional methods by effectively handling the system's nonlinear dynamics. Benefiting from its twin-critic architecture, the TD3 controller exhibited smoother torque profiles and fewer oscillations than the classical LQR. During curved-path obstacle navigation, the TD3 algorithm successfully reduced the root mean square error (RMSE) of the roll angle to 1.23 degrees. These findings confirm that the proposed DRL-based framework provides a highly efficient and adaptable solution for the control and autonomous navigation of self-balancing two-wheeled vehicles.

Keywords: Balance Control, Self-Balancing Bicycle, Behavior-Based Navigation Algorithm, Reaction Wheel, Reinforcement Learning.

۱. مقدمه

پیشرفت فناوری در زمینه‌ی سامانه‌های خودران و هوشمند، پژوهش در حوزه‌ی وسایل نقلیه‌ی دوچرخ خودپایدار را در سال‌های اخیر به یکی از موضوعات پررنگ مهندسی کنترل و مکترونیک تبدیل کرده است. این وسایل به دلیل ابعاد کوچک، جرم پایین، بازدهی بالا و قابلیت مانور زیاد، گزینه‌ای مناسب برای سامانه‌های حمل و نقل سبک شهری محسوب می‌شوند [1]. با این حال، چالش اصلی در طراحی چنین وسایلی، حفظ پایداری در سرعت‌های پایین یا حالت سکون است؛ زیرا در این شرایط، اثرات ژيروسکوپی کاهش یافته و کوچک‌ترین اغتشاش محیطی یا تغییر در مرکز جرم می‌تواند سبب واژگونی وسیله شود. در نتیجه، طراحی سامانه‌های کنترلی دقیق و هوشمند برای تضمین تعادل و حرکت پایدار وسایل دوچرخ از اهمیت ویژه‌ای برخوردار است.

یکی از پرکاربردترین روش‌ها برای حفظ تعادل در سیستم‌های دوچرخ، استفاده از چرخ عکس‌العمل^۱ است که با اعمال گشتاور مخالف به بدنه، سیستم را به وضعیت قائم بازمی‌گرداند. در این راستا، کنداش و کاپیتانی در سال ۲۰۲۳، با طراحی یک مدل دینامیکی دقیق از دوچرخ خودپایدار، نشان دادند که کنترل گشتاور لختی می‌تواند انحراف زاویه‌ی غلتش را تا حدود $\pm 1.5^\circ$ درجه کاهش دهد و زمان تثبیت سیستم را بهبود بخشد [1]. بادگوجار و موهیت در سال ۲۰۲۳ نیز با توسعه‌ی مکانیزم ژيروسکوپ کنترل لحظه‌ای^۲ برای دوچرخ خودپایدار، ساختاری ارائه کردند که توانست نوسانات سیستم را تا ۴۰٪ کاهش دهد و پایداری را در حالت سکون حفظ کند [2]. همچنین، سیلر آلکالا و همکاران در سال ۲۰۲۴ با استفاده از چرخ لختی برای پایداری دوچرخه، نشان دادند که افزایش سرعت پاسخ چرخ عکس‌العمل، پایداری کلی سیستم را در حالت اغتشاش‌های تصادفی به میزان ۲۵٪ بهبود می‌دهد. از سوی دیگر، تران و همکاران (۲۰۲۴) از کنترل‌کننده‌ی تناسبی-انترگالی-مشتقی^۳ برای کنترل دوچرخه‌ی مجهز به چرخ عکس‌العمل استفاده کردند و توانستند با تنظیم بهره‌ها، زاویه‌ی غلتش را در کمتر از ۲.۲ درجه ثابت نگه دارند [3]. روش آن‌ها نشان داد که کنترلرهای کلاسیک،

اگرچه ساده و سریع‌اند، اما در حضور اغتشاش‌های پیچیده یا غیرخطی دقت کافی ندارند. در مطالعه‌ای دیگر، چپو و همکاران در سال ۲۰۲۰ با طراحی یک کنترل‌کننده‌ی هوشمند مقاوم برای دوچرخه‌ی رباطیکی، نشان دادند که ترکیب کنترل تطبیقی و مدل مقاوم می‌تواند خطای زاویه‌ی غلتش را تا ۶۰٪ کاهش دهد [4]. در پژوهش تسبیچا و همکاران در سال ۲۰۲۲، نیز یک استراتژی کنترل مقاوم برای دوچرخه رباطیک ارائه شد که در سرعت‌های پایین از واژگونی جلوگیری می‌کرد [5]. این پژوهش‌ها پایه‌ی قابل توجهی برای توسعه‌ی کنترل‌کننده‌های مقاوم و هوشمند ایجاد کردند.

پژوهش‌های مختلفی نیز بر اساس مدل‌سازی دینامیکی و کنترل مکانیکی انجام شده است. سیکائو و همکاران در سال ۲۰۲۰ با استفاده از مدل پاندول معکوس و فرمان‌پذیری فعال، دوچرخه‌ای طراحی کردند که در مسیر مستقیم قادر به حفظ تعادل بود [6]. هوروب و همکاران در سال ۲۰۲۳ نیز یک سامانه‌ی فن مجرای الکتریکی برای دوچرخه طراحی کردند تا با کنترل جریان هوا بتوانند نیروی مخالف گشتاور واژگونی تولید کند [7]. همچنین کچایا و همکاران در سال ۲۰۲۴ با طراحی سامانه‌ی کنترل خودکار دوچرخه، نشان دادند که استفاده از کنترل مبتنی بر مسیر می‌تواند پایداری دقت ناوبری را هم‌زمان افزایش دهد [8].

در سال‌های اخیر، با گسترش هوش مصنوعی و یادگیری ماشین، استفاده از الگوریتم‌های یادگیری تقویتی^۴ در کنترل سیستم‌های غیرخطی مورد توجه گسترده قرار گرفته است. در این رویکرد، عامل یادگیرنده از طریق تعامل با محیط و دریافت پاداش متناسب با عملکرد، سیاست کنترلی بهینه را بدون نیاز به مدل دقیق سیستم می‌آموزد. زنگ و همکاران در سال ۲۰۲۳، الگوریتم یادگیری تقویتی را برای کنترل ربات‌های دوچرخ در مسیرهای باریک به کار گرفتند و نشان دادند که سیستم قادر است با وجود محدودیت فضا و اغتشاش، تعادل خود را حفظ کند [9]. در ادامه، ژو و همکاران در سال ۲۰۲۲ با ترکیب یادگیری تقویتی و کنترل لغزشی انتهای تطبیقی، مدلی برای کنترل دوچرخه در مسیرهای منحنی و ناهموار ارائه کردند و گزارش دادند که ریشه میانگین مربعات خطا زاویه‌ی غلتش تا ۵۰٪ کاهش یافته است [10].

^۲ PID Controller

^۴ Reinforcement Learning

^۱ Reaction Wheel

^۲ CMG

در همین راستا، آقای جلالی و همکاران در سال ۲۰۲۶ رویکردی ترکیبی نوین مبتنی بر یادگیری تقویتی و کنترل مد لغزشی را برای پایدارسازی سیستم آونگ معکوس بهبودیافته روی گاری که از منظر دینامیکی با سیستم‌های دوچرخ خودپایدار اشتراک ساختاری دارد؛ ارائه کردند. در این پژوهش، کنترل‌کننده مد لغزشی مقاومت لازم در برابر عدم قطعیت‌های مدل و دینامیک غیرخطی را تأمین می‌کند، در حالی که مؤلفه‌ی یادگیری تقویتی به‌عنوان لایه‌ی تطبیقی، دینامیک مدل‌نشده و انحرافات زاویه‌ای بزرگ را جبران می‌سازد. صحت این کنترل‌کننده از طریق تحلیل لیاپانوف اثبات شد و با ده‌ها بار شبیه‌سازی تحت شرایط تصادفی اعتبارسنجی گردید؛ نتایج نشان داد کنترل‌کننده ترکیبی پیشنهادی در ناحیه‌ی غیرخطی به‌طور قابل توجهی از روش‌های کلاسیک مانند تنظیم‌کننده‌ی خطی مربعی و کنترل پیش‌بین مدل، پیشی می‌گیرد. این یافته‌ها نشان می‌دهند که ادغام روش‌های یادگیری تقویتی با کنترل‌کننده‌های مقاوم کلاسیک در پایدارسازی سیستم‌های زیرعملگر غیرخطی مزیت چشم‌گیری دارد [11]. ویرر و همکاران در سال ۲۰۲۴ نیز با بهره‌گیری از یادگیری تقویتی در مدل دوچرخه، الگوریتمی طراحی کردند که قادر بود هم‌زمان پایداری و پیروی مسیر را تضمین کند. در این پژوهش، عامل یادگیرنده در محیطی با موانع متحرک آموزش دید و موفق شد مسیر هدف را با دقت ۹۷٪ دنبال کند. در مطالعه‌ای دیگر، لاکشمی و مانی‌موجی در سال ۲۰۲۴، از الگوریتم گرادیان‌های خط مشی قطعی^۱ برای کنترل دوچرخه خودپایدار استفاده کردند [12]. نتایج نشان داد که این روش زاویه‌ی غلتش را از ۲.۶ درجه به ۱.۲ درجه کاهش داده و زمان همگرایی سیستم را تا ۳۵٪ کاهش داده است. این پژوهش پایه‌ای برای استفاده از روش‌های یادگیری عمیق در کنترل پیوسته‌ی وسایل خودپایدار به شمار می‌رود. افزون بر این، ژانگ و همکاران در سال ۲۰۲۴، استراتژی اجتناب از برخورد مبتنی بر یادگیری تقویتی را برای خودروهای خودران در تقاطع‌های بدون سیگنال ارائه کردند [13]. هرچند این پژوهش مستقیماً درباره‌ی دوچرخه‌های خودپایدار نیست، اما منطق تصمیم‌گیری مبتنی بر پاداش و یادگیری در محیط‌های پویا را تقویت می‌کند. لای و همکاران در سال ۲۰۲۴ نیز با به‌کارگیری کنترل

پیش‌نگر مسیر در دوچرخه‌های هوشمند خودپایدار، نشان دادند که ترکیب پیش‌بینی مسیر و فرمان‌پذیری تطبیقی می‌تواند دقت حرکت را تا ۲۲٪ افزایش دهد [14].

در حوزه‌ی کنترل مقاوم و کاهش مرتبه‌ی مدل، ترونک و همکاران در سال ۲۰۲۵ الگوریتمی ارائه کردند که ضمن حفظ دقت، پیچیدگی محاسباتی کنترل سیستم دوچرخه‌ی خودپایدار را کاهش می‌دهد [15]. در مطالعه‌ای دیگر، دیوریس و همکاران در سال ۲۰۲۵، مدل پایداری دوچرخه‌های باری کم‌سرعت را تحلیل کرد و نشان داد که تنظیم جرم و موقعیت مرکز ثقل نقش اساسی در محدوده‌ی پایداری دارد [16]. این پژوهش‌ها مبنایی برای درک بهتر پویایی وسایل دوچرخ در کاربردهای واقعی فراهم کردند.

پژوهش‌های بادگوجار و موهیت [2] و هارتونو و همکاران [17]، نیز از ترکیب مکانیزم چرخ عکس‌العمل و طراحی مکانیکی بهره گرفتند تا مدل‌های فیزیکی واقعی را بسازند. آن‌ها نشان دادند که در صورت تنظیم بهینه‌ی گشتاور عکس‌العمل، سیستم قادر است در کمتر از ۳ ثانیه از زاویه‌ی انحراف ۱۰ درجه به تعادل بازگردد. دامیان واسیلفسکی در سال ۲۰۲۳، نیز با ساخت یک مدل فیزیکی دوچرخه‌ی خودپایدار، عملکرد کنترل‌کننده‌های سخت‌افزاری را در شرایط واقعی بررسی کردند [18].

از منظر کنترل مسیر و ناوبری، تران و همکاران در سال ۲۰۲۴ از کنترل‌کننده‌ی تناسبی-انتگرالی مشتقی برای کنترل زاویه‌ی فرمان دوچرخه بهره بردند و گزارش کردند که خطای مسیریابی در حضور اغتشاش کمتر از ۳٪ باقی مانده است [3]. همچنین، تین و همکاران در سال ۲۰۲۰ نشان دادند که مدل‌سازی دقیق تعامل بین محورهای مختلف در وسایل دوچرخ می‌تواند عملکرد کنترل‌کننده‌های تناسبی-انتگرالی مشتقی و کنترل‌کننده‌ی خطی درجه دوم^۲ را در مسیرهای منحنی بهبود دهد [19]. در پژوهش دیگری، وانگ و همکاران در سال ۲۰۲۵ با طراحی مدل تک‌چرخ مبتنی بر کنترل‌کننده LQR، پایداری سیستم را در حالت خطی تضمین کرد [20].

وسایل نقلیه‌ی دوچرخ مانند دوچرخه و دوچرخه، به‌دلیل پایداری محدود و ساختار باریک خود، برای حرکت خودکار باید هم‌زمان از موانع محیطی اجتناب کرده و مسیر هدف را دنبال کنند. طراحی

^۲ Linear Quadratic Regulator (LQR)

^۱ Deep Deterministic Policy Gradient (DDPG)

الگوریتم‌هایی که بتوانند ناوبری دقیق را با حفظ تعادل دینامیکی ترکیب کنند، از چالش‌های اصلی در حوزه‌ی وسایل نقلیه‌ی دوچرخ خودران است. در همین راستا، آقای وی و همکاران در سال ۲۰۰۶، مدلی دینامیکی بهبودیافته برای دوچرخه خودران ارائه کردند که در آن اثرات هندسی سیستم فرمان شامل زاویه‌ی میل محور فرمان^۱ و فاصله‌ی پیش‌روی محور فرمان^۲ لحاظ شده است [21]. آنان یک الگوریتم کنترل غیرخطی با ساختار داخلی- خارجی توسعه دادند که هم‌زمان توانایی تعقیب مسیر مرجع و حفظ تعادل وسیله را فراهم می‌سازد. مسیر مطلوب با ترکیب داده‌های سامانه موقعیت‌یاب جهانی^۳ و بینایی ماشین تولید شد و کنترل‌کننده طراحی‌شده، با تضمین پایداری سیستم، امکان حرکت پایدار به سمت هدف را فراهم کرد. نتایج شبیه‌سازی نشان داد که خطای مسیر کمتر از یک متر بوده و سیستم در برابر نویز حسگرها عملکردی پایدار دارد [21]. نوآوری اصلی پژوهش در ترکیب هم‌زمان کنترل تعادل و ناوبری در قالب یک مدل غیرخطی دقیق است که مبنایی مؤثر برای توسعه‌ی سامانه‌های ناوبری هوشمند وسایل نقلیه‌ی دوچرخ محسوب می‌شود [21]. همچنین آقای سینگ و همکاران در سال ۲۰۱۸ مدلی دینامیکی از دوچرخه ارائه کردند و با استفاده از کنترل‌کننده‌ی غالب^۴، روشی برای اجتناب از موانع^۵ پیشنهاد دادند. در این پژوهش، مسیر حرکت مطلوب با ترکیب سه الگوریتم شامل دنبال کردن خط^۶، الگوریتم مماسی برای اجتناب از مانع^۷ و الگوریتم دنبال کردن دیوار^۸ تعیین شد تا دوچرخه بتواند به‌صورت خودکار از موانع ثابت دوری کرده و به سمت هدف حرکت کند [22]. سپس مسیر حاصل به مدل دینامیکی دوچرخه که با استفاده از روش نمودار پیوندی^۹ مدل‌سازی شده بود، اعمال گردید تا سیگنال‌های کنترلی مناسب تولید شود. نتایج شبیه‌سازی در نرم‌افزار متلب^{۱۰} و سیمولینک^{۱۱} نشان داد که سامانه پیشنهادی قادر است با دقت بالا و پایداری مناسب مسیر را دنبال کرده و از موانع اجتناب نماید. نوآوری اصلی این پژوهش در ترکیب الگوریتم‌های ناوبری هوشمند^{۱۲} با کنترل‌کننده‌ی غالب برای

مدل دینامیکی واقعی دوچرخه است که موجب بهبود دقت و پایداری حرکتی نسبت به روش‌های کلاسیک گردیده است [22]. کنار رویکردهای مبتنی بر یادگیری، روش‌های انرژی‌مبنا و لاگرانژین کنترل‌شده نیز از جمله رویکردهای نظری قدرتمند در پایداری وسایل دوچرخ محسوب می‌شوند. تیفی و گنگ در سال ۲۰۱۸، یک چارچوب هندسی نوین برای کنترل هم‌زمان تعادل و حرکت مسطح ربات‌های چرخ‌دار خودپایدار ارائه کردند. در این روش، انرژی جنبشی سیستم به‌گونه‌ای شکل‌دهی می‌شود که جهت قائم بدنه‌ی ربات به یک نقطه‌ی تعادل پایدار غیرخطی برای لاگرانژین کنترل‌شده تبدیل گردد. برای کنترل حرکت مسطح نیز موقعیت و جهت‌گیری ربات به‌عنوان عضوی از گروه اقلیدسی خاص^{۱۳} در فضای دو بعدی (SE(2)) مدل‌سازی شده و از کنترل بازخورد لگاریتمی مبتنی بر مختصات نمایی گروه لی بهره گرفته می‌شود [23].

افزون بر روش‌های ذکر شده، پژوهش‌های متعددی در زمینه کنترل مسیر و ناوبری سیستم‌های رباتیک با استفاده از روش‌های کنترل غیرخطی انجام شده است. به عنوان مثال، استفاده از کنترل‌کننده‌های مقاوم نظیر روش فازی مد لغزشی برای هدایت ربات‌های پرنده در محیط‌های دارای عدم قطعیت نتایج قابل‌قبولی به همراه داشته است [24]. همچنین روش‌های نوین ناوبری و نقشه‌برداری در ربات‌های متحرک نیز به تفصیل مورد ارزیابی قرار گرفته است [25]. با این وجود، در سیستم‌های به شدت غیرخطی مانند دوچرخه‌های خودپایدار که نیازمند حفظ تعادل جانبی در سرعت‌های پایین و هم‌زمان ناوبری در حضور موانع هستند، رویکردهای بدون‌مدل^{۱۴} نظیر الگوریتم‌های یادگیری تقویتی عمیق می‌توانند تطبیق‌پذیری و عملکرد بهتری نسبت به روش‌های کلاسیک یا فازی ارائه دهند.

در مقایسه با این روش، رویکرد پیشنهادی در پژوهش حاضر مبتنی بر یادگیری تقویتی عمیق (TD3/DDPG) است. مزیت اصلی این رویکرد در آن است که برخلاف روش لاگرانژین کنترل‌شده که

^۸ Wall Following Algorithm

^۹ Bond Graph Method

^{۱۰} MATLAB

^{۱۱} Simulink

^{۱۲} Intelligent Navigation Algorithms

^{۱۳} Special Euclidean Group

^{۱۴} Model-free

^۱ Caster Angle

^۲ Trail

^۳ Global Positioning System (GPS)

^۴ Overwhelming Controller

^۵ Obstacle Avoidance

^۶ Line Following

^۷ Tangent Bug Algorithm

بهینه چرخ عکس‌العمل برای حفظ تعادل جانبی و دینامیکی وسیله را بر عهده دارند. این چارچوب چندلایه به گونه‌ای طراحی شده که پایداری، دقت و سازگاری دوچرخه خودپایدار را بهبود می‌بخشد و بستری مناسب برای توسعه وسایل نقلیه خودران سبک و پایدار در محیط‌های واقعی فراهم می‌کند.

معماری کلی چارچوب کنترلی پیشنهادی در شکل ۱ نشان داده شده است. این سیستم از یک ساختار دو حلقه‌ای تشکیل شده است؛ حلقه‌ی داخلی مبتنی بر یادگیری تقویتی که زاویه‌ی غلتش θ_{bike} ، نرخ تغییر زاویه‌ی غلتش $(\dot{\theta}_{bike})$ ، گشتاور چرخ عکس‌العمل $(\tau_{React}(t))$ و سرعت زاویه‌ای آن (ω_{React}) را به عنوان مشاهده دریافت کرده و دستور گشتاور بهینه $(\tau_{React}(t+1))$ را برای حفظ تعادل جانبی دوچرخه تولید می‌کند؛ و حلقه‌ی خارجی ناوبری که در آن الگوریتم ناوبری موقعیت دوچرخه $((x,y)_{bike})$ را از حسگر نصب‌شده دریافت کرده و با استفاده از مختصات هدف $((x,y)_{Goal})$ و موانع $((x,y)_{Obs})$ که از پیش تعریف شده‌اند، زاویه‌ی فرمان مناسب را از طریق کنترل‌کننده‌ی PID محاسبه می‌کند. همچنین، الگوریتم ناوبری سرعت خطی دوچرخه را نیز تنظیم می‌کند؛ به‌طوری‌که در تمام مسیر، از جمله هنگام اجتناب از موانع، سرعت ثابت ۳ متر بر ثانیه حفظ می‌شود و تنها پس از رسیدن به نقطه‌ی هدف، سرعت به صفر کاهش می‌یابد.

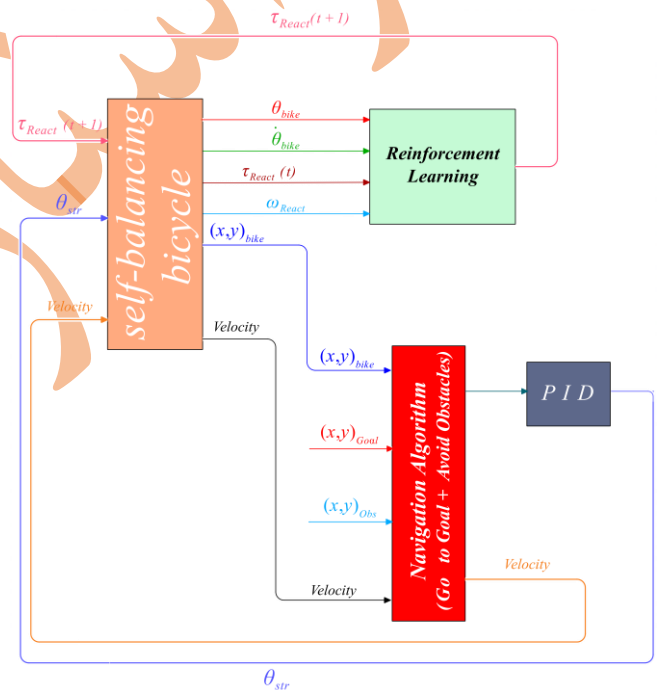
۲. مدل‌سازی دوچرخه

مدل دینامیکی دوچرخه خودپایدار بر پایه مدل ساده‌شده دوچرخه‌ای توسعه یافته است. این انتخاب به دلیل شباهت در درجات آزادی اصلی میان دوچرخه و دوچرخه و سهولت تحلیل دینامیکی انجام شد. تفاوت اصلی مدل پیشنهادی با مدل متعارف، افزودن چرخ عکس‌العمل در بخش میانی بدنه است که به عنوان عملگر لختی، گشتاور جانبی لازم برای حفظ تعادل وسیله را ایجاد می‌کند. از دید دینامیکی، این ساختار مشابه یک پاندول معکوس با عملگر لختی عمل کرده و پایه مناسبی برای طراحی کنترل‌کننده‌های پایداری فراهم می‌کند.

مدل سه‌بعدی سیستم در نرم‌افزار سالدورکس طراحی شد و اجزای اصلی شامل بدنه، چرخ‌های جلو و عقب و چرخ عکس‌العمل مدل‌سازی گردیدند. اجزای کم‌تأثیر مانند پدال و زنجیر برای کاهش

نیازمند مدل دینامیکی دقیق و استخراج تحلیلی معادلات انرژی است، الگوریتم‌های یادگیری تقویتی بدون نیاز به مدل صریح از سیستم، سیاست کنترلی بهینه را از طریق تعامل با محیط شبیه‌سازی می‌آموزند. این ویژگی، انعطاف‌پذیری بیشتری در مواجهه با غیرخطی بودن دینامیک، اغتشاشات محیطی و تغییرات پارامتری فراهم می‌سازد.

روش‌های کنترل کلاسیک مانند PID و LQR برای ناوبری عملکرد قابل قبول دارند، اما در مواجهه با اغتشاشات غیرخطی و تغییرات دینامیکی محدودیت پیدا می‌کنند. الگوریتم‌های یادگیری تقویتی مانند خط‌مشی قطعی عمیق و گرادیان خط‌مشی قطعی دوگانه‌ی تأخیری قادر به یادگیری سیاست‌های پایدار و بهینه هستند، اما به داده و زمان همگرایی بیشتری نیاز دارند.



شکل ۱. نمودار بلوکی کلی چارچوب کنترلی دو حلقه‌ای پیشنهادی برای دوچرخه خودپایدار ساز

Figure 1. Overall block diagram of the proposed dual-loop control framework for the self-balancing bicycle

در این پژوهش، یک چارچوب ترکیبی و هوشمند ارائه شده است که هم‌زمان از مزایای کنترل کلاسیک و یادگیری تقویتی بهره می‌برد. در این ساختار، کنترل‌کننده PID مسئول تنظیم زاویه فرمان و اجتناب از موانع برای حفظ مسیر دوچرخه در محیط‌های پیچیده است، در حالی که الگوریتم‌های یادگیری تقویتی وظیفه تولید گشتاور

پیچیدگی محاسباتی حذف شدند. مدل نهایی با جرم ۷ کیلوگرم و طول ۱۲۰ سانتی متر دارای چهار درجه آزادی حرکتی (زاویه غلتش، زاویه فرمان، جابجایی طولی و عرضی) و سه درجه آزادی فعال شامل گشتاور چرخ عقب، زاویه فرمان و گشتاور چرخ عکس العمل است. مدل دینامیکی به صورت یک سامانه‌ی سه‌ورودی-چهارخروجی تعریف می‌شود. ورودی‌های کنترلی شامل سرعت خطی (v_{rear})، زاویه‌ی فرمان (θ_{str}) و گشتاور چرخ عکس العمل (τ_{react}) بوده و خروجی‌ها عبارتند از زاویه‌ی غلتش (ϕ)، نرخ تغییر زاویه‌ی غلتش ($\dot{\phi}$)، گشتاور چرخ عکس العمل (τ_{react}) و سرعت زاویه‌ای چرخ عکس العمل (ω_{react}).



شکل ۲. مدل سه‌بعدی دوچرخه خودپایدارساز

Figure 2. Three-dimensional model of the self-stabilizing bicycle
مدل طراحی شده پس از تکمیل در نرم‌افزار سالدورکس از طریق رابط نرم‌افزاری به محیط شبیه‌سازی چندجسمی سیمولینک سیم‌اسکیپ^۱ در متلب منتقل شد. در این فرآیند، تمامی مختصات هندسی، قیود حرکتی و اتصالات صلب به صورت خودکار وارد محیط شبیه‌سازی گردید تا انطباق کامل میان مدل هندسی و دینامیکی برقرار شود. این بستر امکان تحلیل دینامیکی دقیق، بررسی پاسخ‌های گذرا و طراحی کنترل‌کننده‌های هوشمند را فراهم می‌سازد. در ادامه، سامانه‌ی کنترل تعادل با استفاده از بازخورد زاویه‌ی غلتش (θ) از حسگر زاویه‌ای طراحی شد. کنترل‌کننده با بهره‌گیری از الگوریتم‌های یادگیری تقویتی، گشتاور بهینه‌ی چرخ عکس العمل را محاسبه کرده و با اعمال گشتاور عکس‌عملی مناسب، پایداری جانبی دوچرخه را در شرایط مختلف حرکتی تضمین می‌کند.

یکی از روش‌های کارآمد برای پایدارسازی دوچرخه‌های خودپایدارساز، استفاده از چرخ عکس‌عمل است. اساس عملکرد این روش بر اصل پایداری تکانه‌ی زاویه‌ای استوار است؛ بدین معنا که اگر یک دیسک لختی با سرعت زاویه‌ای بالا حول محور خود بچرخد،

تغییر سرعت یا تغییر جهت چرخش آن باعث ایجاد گشتاور عکس‌عملی بر روی بدنه‌ی وسیله می‌شود. این گشتاور به صورت مستقیم بر زاویه‌ی غلتش دوچرخه اثر گذاشته و می‌تواند انحراف آن از وضعیت قائم را اصلاح کند. با افزایش یا کاهش سرعت زاویه‌ای چرخ عکس‌عمل که توسط موتور کنترل می‌شود، گشتاور عکس‌عملی متناسب تولید می‌گردد که مخالف انحراف جانبی وسیله عمل کرده و آن را به حالت تعادل بازمی‌گرداند.

مدل‌سازی دینامیکی دوچرخه مجهز به چرخ عکس‌عمل در این پژوهش بر پایه معادلات و مدل ارائه‌شده در [1] توسعه یافته است. در این ساختار، پایداری سامانه بر اساس اثر ژيروسکوپی تحلیل می‌شود و از روابط تکانه (معادله ۱) و گشتاور پیشروی (رابطه ۲) برای بیان اندرکنش بین چرخ عکس‌عمل و بدنه استفاده می‌گردد.

$$H = \omega I \quad 1$$

$$\tau = H\Omega \quad 2$$

این بیان ریاضی امکان بررسی نقش ابرمولفه‌هایی مانند جرم، شعاع و سرعت زاویه‌ای چرخ را در تولید گشتاور مقاومتی و بازگرداندن سیستم به وضعیت قائم فراهم می‌سازد.

در رویکرد دوم، استخراج معادلات حرکت (معادلات ۳ تا ۱۰) بر پایه رویکرد لاگرانژی ارائه‌شده در [23] توسعه یافته است. در این روش، ابتدا انرژی جنبشی و انرژی پتانسیل اجزای سیستم محاسبه شده و سپس لاگرانژ کل به صورت معادله ۳ تعریف می‌گردد.

$$L = K - V \quad 3$$

با به‌کارگیری معادله اوایلر-لاگرانژ (معادله ۴) می‌توان معادلات دینامیکی سیستم را استخراج نمود.

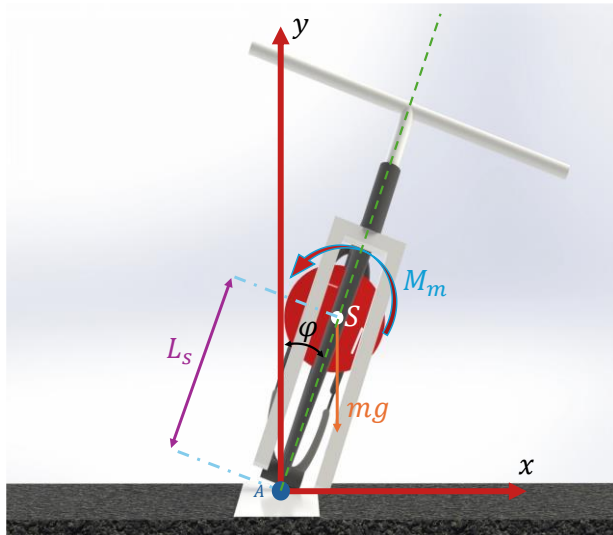
$$\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) - \frac{\partial L}{\partial q_i} = \tau_i, \quad i = 1, 2, 3, \dots \quad 4$$

با توجه به معادله ۴، مدل دینامیکی با استفاده از معادله‌ی ضربه تکانه برای یک جسم صلب در حرکت صفحه‌ای، حول محور ثابت Z در نقطه‌ی S استخراج می‌گردد.

$$-J_a \ddot{\phi} = -mgL_s \sin(\phi) + M_m \quad 5$$

با خطی‌سازی در نواحی زاویه‌های کوچک، رابطه ۵ به فرم ساده‌تر درمی‌آید.

^۱ SimScape



شکل ۳. مدل فیزیکی و دستگاه مختصات دوچرخه خودپایدارساز با چرخ عکس‌العمل در صفحه غلتش

Figure 3. Physical model and coordinate system of the self-balancing bicycle with a reaction wheel in the roll plane

Table 1. Parameters used in the above equations

جدول ۱. پارامترهای مورد استفاده در معادلات بالا

No.	Symbol	Description
1	J_a	Moment of inertia
2	$\ddot{\phi}$	Roll angular acceleration
3	m	Motorcycle mass
4	ϕ	Roll angle
5	M_m	Torque applied to the reaction wheel
6	L_s	Distance between the center of mass and the ground
7	x	System state vector
8	A	System dynamic matrix
9	B	Actuator input matrix (vector)
10	C	Sensor output (input) matrix
11	ϕ_{max}	Maximum roll angle
12	M_m^{max}	Maximum applied torque

$$\ddot{\phi} \frac{mgL_s}{J_a} = -\frac{M_m}{J_a} \quad 6$$

با توجه به معادله‌ی خطی‌سازی‌شده (رابطه ۶)، سیستم را می‌توان به صورت مدل فضای حالت بازنویسی کرد. بردار حالت شامل زاویه‌ی غلتش (ϕ)، نرخ تغییر زاویه‌ی غلتش ($\dot{\phi}$)، گشتاور موتور (M_m) و سرعت زاویه‌ای چرخ عکس‌العمل (ω_m) خواهد بود.

$$\dot{x} = Ax + Bu \quad 7$$

$$x = [\phi \quad \dot{\phi} \quad M_m \quad \omega_m]^T \quad 8$$

$$u = M_m \quad 9$$

ماتریس‌های A و B بر اساس ابرمولفه‌های فیزیکی وسیله شامل جرم، ممان اینرسی بدنه، فاصله‌ی مرکز جرم از زمین، ضرایب الکتریکی موتور و لختی دیسک تعیین می‌شوند. این فرم ریاضی امکان استفاده از روش‌های کنترلی پیشرفته همچون کنترل‌کننده بهینه LQR و یا کنترل‌کننده PID برای پایدارسازی دوچرخه را فراهم می‌کند.

همچنین باید توجه داشت که مقدار گشتاور تولیدی چرخ عکس‌العمل محدود است و از یک حد آستانه بیشتر نمی‌تواند تعادل وسیله را حفظ کند. این حد با توجه به ابرمولفه‌های سیستم از رابطه‌ی زیر به دست می‌آید.

$$\phi_{max} = \frac{M_m^{max}}{mgL_s} \quad 10$$

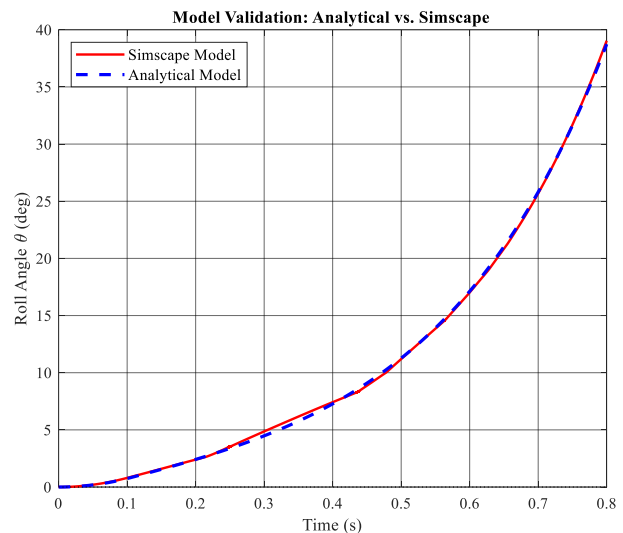
این رابطه نشان می‌دهد که اگر زاویه‌ی غلتش بیشتر از ϕ_{max} شود، حتی با پیشینه‌ی توان موتور نیز امکان بازگرداندن وسیله به حالت تعادل وجود نخواهد داشت. بنابراین طراحی سامانه باید به گونه‌ای باشد که انحراف‌های زاویه‌ای کوچک و متوسط به سرعت اصلاح شوند تا از رسیدن به محدوده‌ی ناپایداری جلوگیری شود. پارامترها و نمادهای به کاررفته در این معادلات در جدول ۱ خلاصه شده‌اند.

فرمان را به سامانه‌ی محرکه ارسال می‌کند. بنابراین، مباحث پیشرفته‌تری نظیر نقشه‌برداری معنایی و هم‌زمان (به‌عنوان نمونه راهکارهای نقشه‌برداری و اکتشاف توسعه‌یافته برای ربات‌های خودگردان در [25])، فیوژن سنسوری و یا تخمین‌گرهای حالت در این معماری لحاظ نشده و مختصات محیطی به‌صورت مفروض در اختیار الگوریتم قرار می‌گیرند.

الگوریتم‌های ناوبری دارای روش‌های گوناگونی هستند که بسته به نوع کاربرد و ساختار سامانه، مورد استفاده قرار می‌گیرند. یکی از رویکردهای پرکاربرد در این زمینه، کنترل رفتار مینا^۱ است. در این روش، رفتارهای حرکتی ساده به‌صورت ماژول‌های مستقل تعریف شده و ترکیب آن‌ها منجر به شکل‌گیری رفتارهای پیچیده‌تر می‌شود. از جمله رفتارهای اصلی در این دسته می‌توان به رفتار رسیدن به هدف^۲، رفتار اجتناب از موانع^۳ و رفتار دنبال کردن دیواره‌ها^۴ اشاره کرد که هسته‌ی اصلی ناوبری هوشمند در وسایل نقلیه خودران را تشکیل می‌دهند.

۱.۱.۲. رسیدن به هدف

رفتار رسیدن به هدف یکی از اساسی‌ترین مؤلفه‌های الگوریتم ناوبری است که در آن عامل تلاش می‌کند از موقعیت مبدأ به سمت مقصد حرکت نماید. راهبرد اصلی این رفتار بر مبنای کاهش خطای زاویه‌ای میان محور حرکت عامل و جهت مقصد تعریف می‌شود. بدین ترتیب، سامانه به‌صورت پیوسته زاویه‌ی بین جهت حرکت و موقعیت هدف را محاسبه کرده و فرمان‌های کنترلی را به‌گونه‌ای تولید می‌کند که این زاویه به حداقل برسد. در نتیجه، عامل به تدریج در راستای هدف قرار گرفته و مسیر خود را به‌صورت پایدار و مستقیم به‌سوی مقصد هموار می‌سازد. هدف نهایی این رفتار، دستیابی به حرکت نرم و هم‌جهت با مقصد تا رسیدن کامل به نقطه‌ی هدف است.



شکل ۴. اعتبارسنجی مدل: مقایسه پاسخ حلقه باز زاویه غلتش بین مدل تحلیلی لاگرانژ و شبیه‌سازی چندجسمی Simscape تحت تحریک گشتاور پالس یکسان

Figure 4. Model validation: open-loop roll angle response comparison between the analytical Lagrangian model and the Simscape multibody simulation under identical pulse torque excitation

به‌منظور اعتبارسنجی مدل ریاضی استخراج‌شده، پاسخ دینامیکی حلقه‌باز مدل خطی‌شده‌ی فضای حالت با شبیه‌سازی چندجسمی سه‌بعدی در محیط سیمولینک مقایسه شد. در این آزمایش، یک ضربه‌ی گشتاوری به میزان ۱ نیوتن‌متر با مدت زمان ۰.۱ ثانیه به چرخ عکس‌العمل در هر دو مدل اعمال گردید و پاسخ زاویه‌ی غلتش دو چرخ ثبت شد. پارامترهای فیزیکی مندرج در جدول ۶ مستقیماً از مدل سه‌بعدی سالی‌دورکس استخراج گردیدند. همان‌طور که در شکل ۴ مشاهده می‌شود، پاسخ‌های دو مدل تطابق بسیار خوبی با یکدیگر دارند که صحت معادلات دینامیکی استخراج‌شده و دقت مدل شبیه‌سازی را تأیید می‌کند.

۱.۲. الگوریتم ناوبری

الگوریتم ناوبری در این پژوهش صرفاً به معنای رهگیری مسیر و کنترل سطح پایین برای رسیدن به نقطه‌ی مقصد و اجتناب از موانع با مختصات مشخص است. در این فرآیند، عامل با دریافت پیوسته‌ی موقعیت خود و مختصات از پیش تعیین‌شده‌ی هدف و موانع، مسیر مناسب را تعیین کرده و فرمان‌های حرکتی نظیر سرعت و زاویه‌ی

^۲ Avoid Obstacle

^۳ Wall Following

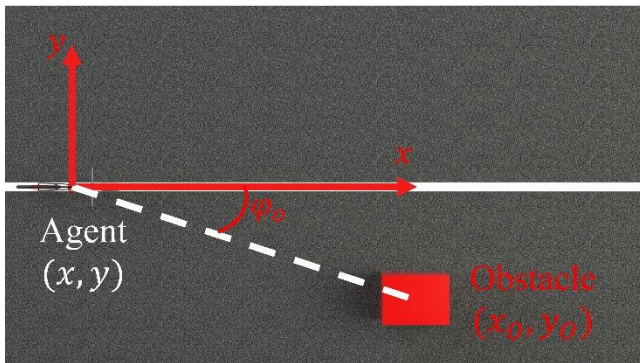
^۱ Behavior-Based Control

^۲ Go to Goal

۲.۱.۲. اجتناب از موانع

در رفتار اجتناب از موانع، عامل با استفاده از داده‌های حسگرها، موانع محیطی را شناسایی کرده و مسیر حرکت خود را به گونه‌ای تنظیم می‌کند که ضمن حفظ جهت کلی حرکت به سمت مقصد، از برخورد با موانع جلوگیری شود. برای پیاده‌سازی این رفتار، از رویکرد برداری مبتنی بر مدل دیسک انتزاعی استفاده شده است. در این مدل، محدوده‌های اطراف دوچرخه به صورت یک ناحیه‌ی دایره‌ای در نظر گرفته می‌شود که شعاع آن متناظر با برد پوشش حسگرهای محیطی است. هر حسگر، حضور مانع را در شعاع مشخصی از موقعیت خود پایش کرده و در صورت تشخیص مانع، اطلاعات مربوط به موقعیت و فاصله را به واحد تصمیم‌گیری ارسال می‌کند. در نتیجه، ناحیه‌ی دیسکی پیرامون وسیله به عنوان حریم ایمنی عمل کرده و ورود هر مانع به این محدوده باعث فعال شدن رفتار اجتناب می‌گردد. استفاده از این مدل موجب ساده‌سازی محاسبات هندسی و افزایش سرعت واکنش سیستم در محیط‌های پویا می‌شود.

$$\varphi_G = \text{atan}2(y_G - y, x_G - x) \quad 13$$



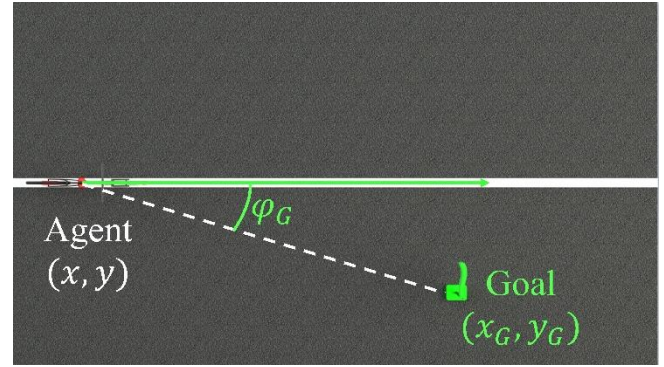
شکل ۶. رفتار اجتناب از مانع

Figure 6. Obstacle avoidance behavior

مطابق شکل ۶، مختصات عامل (x, y) و مختصات مانع (x_O, y_O) در دستگاه مرجع تعریف شده‌اند. بردار واصل از عامل به مانع به صورت زیر بیان می‌شود.

$$\mathbf{r}_{r_O} = \begin{bmatrix} x_O - x \\ y_O - y \end{bmatrix} \quad 14$$

زاویه مانع نسبت به جهت حرکت عامل (φ_O) با استفاده از رابطه ۱۴ به دست می‌آید. همان‌طور که در روش رسیدن به هدف توضیح داده شد، رابطه ۱۵ به دلیل پرش زاویه در عبور از محور مختصات به رابطه ۱۶ تبدیل می‌شود.



شکل ۵. رفتار حرکت به سمت هدف

Figure 5. Goal to goal behavior

برای تعیین زاویه‌ی φ_G که همان زاویه‌ی بین جهت حرکت عامل و خط واصل عامل تا هدف است، ابتدا بردارهای مربوط به وضعیت عامل و هدف تعریف می‌شوند (شکل ۵). فرض کنیم مختصات عامل (x, y) و مختصات هدف (x_G, y_G) باشد. در این صورت، بردار خط واصل عامل به هدف با رابطه ۱۱ بدست می‌آید.

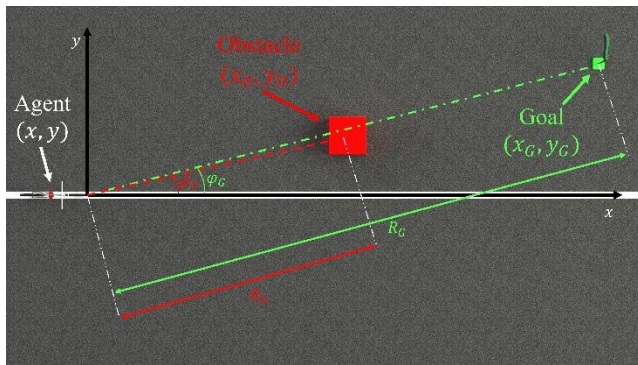
$$\mathbf{r}_{r_G} = \begin{bmatrix} x_G - x \\ y_G - y \end{bmatrix} \quad 11$$

جهت‌گیری عامل نسبت به محور افق با زاویه‌ی θ مشخص می‌شود. بر این اساس، زاویه‌ی مطلوب φ_G از رابطه‌ی ۱۲ به دست می‌آید.

$$\varphi_G = \tan^{-1} \left(\frac{x_G - x}{y_G - y} \right) \quad 12$$

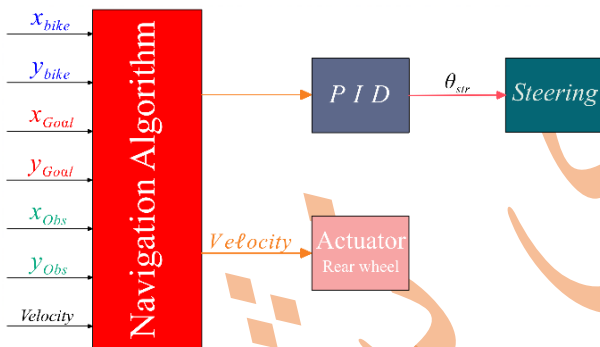
از آن‌جا که تابع $\tan^{-1}$ تنها زاویه را در بازه‌ی $\pm 90^\circ$ باز می‌گرداند و قادر به تشخیص ربع صحیح مختصات نیست، هنگام عبور بردار از محور x منفی ناپیوستگی زاویه رخ می‌دهد. برای رفع این مشکل، از تابع دو ورودی $\text{atan}2(y, x)$ استفاده شد که با در نظر گرفتن علامت هر دو مؤلفه، موقعیت بردار را در ربع صحیح تعیین کرده و زاویه‌ی پیوسته و دقیقی فراهم می‌سازد. این ویژگی موجب بهبود پایداری الگوریتم ناوبری می‌شود. برای هدایت وسیله به مسیر مطلوب، تابع خطا به صورت ترکیبی از خطای زاویه‌ای و مکانی تعریف گردید تا میزان انحراف از مسیر مرجع را نشان دهد. این تابع به عنوان ورودی کنترل‌کننده‌ی PID به کار رفت تا زاویه‌ی فرمان مناسب را تولید کند. در نتیجه، جهت‌گیری وسیله تدریجاً با مسیر هدف هم‌راستا شده و فاصله مکانی کاهش می‌یابد. در نزدیکی مقصد، خطاها به صفر میل کرده و کنترل‌کننده با کاهش خروجی خود، سامانه را در نقطه‌ی هدف به حالت تعادل پایدار می‌رساند.

می‌یابد، و در صورت غالب شدن رفتار اجتناب از موانع، احتمال توقف یا انحراف از مسیر اصلی افزایش پیدا می‌کند. روش دیگر، استفاده از مکانیزم اولویت‌بندی^۲ است؛ به گونه‌ای که در حالت عادی حرکت به سمت مقصد انجام می‌شود، اما در هنگام نزدیک شدن به موانع، رفتار اجتناب از مانع بر سایر رفتارها اولویت یافته و پس از رفع خطر، حرکت مستقیم به هدف از سر گرفته می‌شود.



شکل ۷. طراحی مفهومی الگوریتم ناوبری با ترکیب رفتارها

Figure 7. Conceptual design of the navigation algorithm combining behaviors



شکل ۸. بلوک دیاگرام الگوریتم ناوبری رفتارمبنی

Figure 8. Block diagram of the behavior-based navigation algorithm

Table 2. Components of equations 11-20

جدول ۲. مؤلفه‌ها و اجزای معادلات ۱۱ تا ۲۰

Symbol	Description
x	Bicycle coordinate in the x direction
y	Bicycle coordinate in the y direction
x_O	Obstacle coordinate in the x direction
y_O	Obstacle coordinate in the y direction
x_G	Target coordinate in the x direction
y_G	Target coordinate in the y direction

$$\varphi_O = \tan^{-1} \left(\frac{x_O - x}{y_O - y} \right) \quad 15$$

$$\varphi_O = \text{atan2}(y_O - y, x_O - x) \quad 16$$

به‌طور کلی، الگوریتم ناوبری را می‌توان مغز تصمیم‌گیری سیستم دانست که با ترکیب رفتارهای هدفمند و حفاظتی، امکان حرکت ایمن، کارآمد و پایدار را در محیط‌های ایستا یا پویا برای عامل هوشمند فراهم می‌آورد. برای افزایش ایمنی حرکت دوطرفه خودپایدار، لازم است سازوکاری برای تشخیص و اجتناب از موانع در مسیر پیش‌بینی شود.

۳.۱.۲. ترکیب دو رفتار

ترکیب رفتارهای رسیدن به هدف و اجتناب از موانع در الگوریتم ناوبری نقش کلیدی دارد. اگر تنها رفتار رسیدن به هدف فعال باشد، احتمال برخورد با موانع بالا خواهد بود، در حالی که اجرای تنها رفتار اجتناب از موانع می‌تواند موجب سرگردانی عامل در محیط و ناتوانی در رسیدن به مقصد شود. از این‌رو، لازم است سازوکاری طراحی شود تا این دو رفتار به‌صورت هم‌زمان و هماهنگ عمل کنند. برای تحقق این هدف، راهکارهای گوناگونی پیشنهاد شده است. یکی از رویکردهای ساده، تعریف قوانین حرکتی مستقل است.

۱. فرار از مانع

$$\varphi_1 = \varphi_O + \pi \quad 17$$

۲. عمود به مانع حرکت کردن

$$\varphi_2 = \varphi_O \pm \frac{\pi}{2} \quad 18$$

۳. نادیده گرفتن مانع و حرکت به سمت هدف

$$\varphi_2 = \varphi_G \quad 19$$

۴. یکی از روش‌های مؤثرتر، ترکیب خروجی هر دو رفتار به‌صورت مجموع وزنی^۱ است. در این روش، فرمان نهایی از ترکیب بردار کنترلی حاصل از رفتار رسیدن به هدف و بردار کنترلی حاصل از اجتناب از مانع محاسبه می‌شود.

$$\varphi_2 = f(\varphi_G, \varphi_O) \quad 20$$

تنظیم ضرایب وزن‌دهی در این ترکیب اهمیت زیادی دارد؛ زیرا چنانچه سهم رفتار رسیدن به هدف بیش از حد باشد، ایمنی کاهش

^۲ Priority Mechanism

^۱ Weighted Summation

نادیده گرفتن هر متغیر کلیدی می‌تواند باعث مدل‌سازی ناقص و ناپایداری سیستم شود.

در حرکت دوچرخه روی مسیر منحنی، علاوه بر حفظ تعادل جانبی، باید تغییر جهت نیز به‌طور دقیق کنترل شود تا از انحراف از مسیر جلوگیری گردد. از آنجا که در این حالت نیروها و تغییرات دینامیکی بیشتری رخ می‌دهد، فضای حالت باید متغیرهای بیشتری را شامل شود تا رفتار دوچرخه در مسیرهای پیچیده به‌درستی توصیف گردد. بنابراین، در مسیرهای منحنی، متغیرهایی مانند زاویه انحراف از مسیر و نرخ تغییر آن نیز به فضای حالت اضافه می‌شوند تا کنترل دقیق‌تر و پایدارتر انجام گیرد.

$$s(t) = [\theta(t), \dot{\theta}(t), \ddot{\theta}(t), u(t), u_{\text{prev}}(t), u_{2\text{prev}}(t), \dot{z}(t), \ddot{z}(t), z(t), \dot{\delta}(t), \ddot{\delta}(t), \delta(t), \dot{v}(t), \ddot{v}(t), v(t), \dot{v}(t)] \quad 21$$

Table 3. Components of the state space

جدول ۳. مؤلفه‌های فضای حالت

Symbol	Description
$\theta_{bike}(t)$	Roll angle of the bicycle
$\dot{\theta}_{bike}(t)$	Roll angular velocity
$\ddot{\theta}_{bike}(t)$	Roll angular acceleration for a more detailed analysis of the roll dynamics
$\omega_{React}(t)$	Reaction wheel angular velocity
$\dot{\omega}_{React}(t)$	Rate of change of the reaction-wheel angular velocity
τ_{React}	Control input, typically the torque applied to the reaction wheel
$\tau_{React\text{prev}}$	Torque at the previous time step
$u_{2\text{prev}}(t)$	Torque at two previous time steps
$z(t)$	Ground-height disturbance
$\dot{z}(t)$	First derivative of ground height
$\ddot{z}(t)$	Second derivative of ground height
$\delta(t)$	Steering angle for guiding the vehicle's path
$\dot{\delta}(t)$	Rate of change of the steering angle
$\ddot{\delta}(t)$	Steering-angle acceleration
$v(t)$	Linear velocity of the bicycle
$\dot{v}(t)$	Linear acceleration of the bicycle

فضای حالت طراحی شده به‌صورت جامع، وضعیت حرکتی دوچرخه را در مسیرهای مستقیم و منحنی توصیف می‌کند. این فضا علاوه بر

φ_O	Instantaneous angle between the bicycle and the obstacle
φ_G	Instantaneous angle between the bicycle and the target
R_G	Vector distance between the bicycle and the target
R_O	Vector distance between the bicycle and the obstacle
Velocity	Linear velocity of the bicycle
θ_{str}	Steering angle
θ_{roll}	Roll angle

۲.۲. یادگیری تقویتی

یادگیری تقویتی روشی نوین در هوش مصنوعی است که به عامل هوشمند اجازه می‌دهد با تعامل مستقیم با محیط و دریافت بازخورد، سیاست کنترلی بهینه برای سامانه‌های پیچیده و غیرخطی بیاموزد. این روش می‌تواند تعادل دینامیکی دوچرخه را در شرایط اغتشاش و مسیرهای ناهموار حفظ کرده و نسبت به کنترل‌کننده‌های کلاسیک از پایداری و سازگاری بالاتری برخوردار است. عامل با مشاهده متغیرهای حالت شامل زاویه غلتش، نرخ تغییر زاویه، سرعت خطی و شتاب جانبی، گشتاور کنترلی مناسب را تولید می‌کند تا تعادل دوچرخه حفظ شود. مزیت اصلی یادگیری تقویتی، توانایی یادگیری مداوم و سازگاری در شرایط ناشناخته است، در حالی که کنترل‌کننده‌های کلاسیک به مدل دقیق سیستم نیاز دارند.

در این پژوهش، یادگیری تقویتی در سیمولینک متلب پیاده‌سازی شده و بلوک عامل دارای سه ورودی (بردار مشاهده، پاداش و سیگنال پایان) و یک خروجی (عمل) است. خروجی عامل به‌صورت گشتاور کنترلی به چرخ عکس‌العمل اعمال می‌شود. برای عملکرد صحیح، مؤلفه‌های اصلی شامل فضای حالت، فضای عمل و تابع پاداش به‌طور دقیق تعریف شده‌اند تا پایه‌ای برای یادگیری و بهینه‌سازی سیاست کنترلی فراهم گردد.

۱.۲.۲. فضای حالت

فضای حالت مجموعه‌ای از متغیرهایی است که وضعیت لحظه‌ای سیستم را توصیف کرده و تمام اطلاعات لازم برای پیش‌بینی رفتار آینده آن را در بر دارد. این فضا پایه اصلی طراحی کنترل‌کننده‌ها و الگوریتم‌های یادگیری تقویتی محسوب می‌شود. در کنترل تعادل دوچرخه، انتخاب دقیق متغیرهای حالت اهمیت زیادی دارد، زیرا

تحلیل دینامیک جانبی، تغییرات جهت حرکت و انحراف مسیر را نیز شامل می‌شود. در حرکت مستقیم، تمرکز بر حفظ تعادل جانبی و پایداری طولی است، در حالی که در مسیر منحنی، متغیرهایی مانند زاویه انحراف و نرخ تغییر جهت به فضای حالت افزوده شده‌اند تا کنترل مسیر دقیق‌تر انجام شود. این ساختار به الگوریتم یادگیری تقویتی امکان می‌دهد سیاست‌های کنترلی بهینه برای حفظ تعادل و هدایت دوچرخه در شرایط مختلف محیطی بیاموزد و عملکرد پایداری در مسیرهای پیچیده ارائه دهد. متغیرهای فضای حالت در جدول ۳ معرفی شده‌اند

۲.۲.۲. تابع پاداش

طراحی تابع پاداش از بخش‌های کلیدی یادگیری تقویتی است، زیرا عامل تنها از طریق سیگنال‌های پاداش می‌تواند رفتار بهینه را بیاموزد. در این پژوهش، تابع پاداشی چندجزئی و ترکیبی طراحی شده که هم‌زمان معیارهایی مانند حفظ تعادل بدنه، بازگشت به حالت قائم، پایداری دینامیکی، ایمنی سخت‌افزاری و نرمی عملکرد عملگر را در بر می‌گیرد. برای هر جزء وزن جداگانه‌ای تعیین شده و پاداش نهایی با ترکیب خطی آن‌ها محاسبه می‌شود. این ساختار باعث می‌شود عامل علاوه بر حفظ تعادل، رفتارهایی مانند جلوگیری از اشباع عملگر، کنترل سرعت زاویه‌ای چرخ عکس‌العمل و اعمال گشتاور نرم و پیوسته را نیز یاد بگیرد. بدین ترتیب، تابع پاداش طراحی شده نقش اصلی در یادگیری سیاست کنترلی پایدار و ایمن ایفا می‌کند.

– پاداش حفظ تعادل در حالت قائم

در این مؤلفه، هدف کنترل نگه‌داشتن زاویه غلتش (θ) در نزدیکی صفر و محدود کردن نرخ زاویه‌ای ($\dot{\theta}$) است. برای این منظور، ناحیه‌ای مطلوب پیرامون مقدار صفر برای هر دو متغیر تعریف شده است. مقدار پاداش در این محدوده با تابع گاوسی مدل‌سازی می‌شود؛ به‌گونه‌ای که هرچه زاویه و نرخ زاویه‌ای به مقدار تعادل نزدیک‌تر باشند، پاداش بیشتر و با افزایش فاصله از نقطه تعادل، پاداش به‌صورت پیوسته کاهش می‌یابد.

$$R_{\text{good}} = \exp\left(-\frac{1}{2}\left(\frac{\theta_{\text{bike}}}{\theta_{\text{well}}}\right)^2\right) \exp\left(-\frac{1}{2}\left(\frac{\dot{\theta}_{\text{bike}}}{\dot{\theta}_{\text{well}}}\right)^2\right) \quad 22$$

– پاداش هم‌ترازی و بازگشت به حالت قائم
به‌منظور افزایش پایداری دینامیکی سیستم، مؤلفه‌ای از تابع پاداش طراحی شد تا عامل را در شرایط اغتشاش تشویق کند که جهت تغییرات متغیرهای حالت به‌گونه‌ای باشد که خطای زاویه‌ای کاهش یافته و سامانه به سمت وضعیت تعادل حرکت کند. این مؤلفه به‌صورت زیر تعریف می‌شود.

$$R_{\text{align}} = \tanh\left(-\frac{\theta_{\text{bike}}}{\theta_{\text{well}}}\right) \quad 23$$

رابطه ۲۳ بیانگر پاداش هم‌ترازی^۱ یا بازگشت به حالت قائم است. این تابع عامل را به انجام اقدامات اصلاحی سریع و مؤثر در جهت کاهش خطای زاویه‌ای و بازگرداندن دوچرخه به وضعیت تعادل در کوتاه‌ترین زمان ممکن ترغیب می‌کند.

– پاداش کاهش انرژی لیاپانوفی

برای تضمین پایداری دینامیکی سیستم، از مفهوم انرژی لیاپانوف استفاده شده است. در این روش، یک تابع انرژی شبه‌لیاپانوف شامل زاویه غلتش، سرعت زاویه‌ای و سرعت زاویه‌ای چرخ عکس‌العمل تعریف گردید. این تابع مطابق رابطه ۲۴ بیان می‌شود.

$$L(t) = a\theta_{\text{bike}}^2 + b\dot{\theta}_{\text{bike}}^2 + c\omega_{\text{React}}^2 \quad 24$$

هدف عامل آن است که مقدار انرژی تعریف‌شده در رابطه ۲۴ به مرور زمان کاهش یابد. بنابراین تغییرات انرژی در دو گام متوالی به‌صورت رابطه ۲۵ در تابع پاداش لحاظ گردید.

$$R_{\text{lyap}} = -\Delta L = L(t) - L(t-1) \quad 25$$

علاوه بر حفظ تعادل، ضروری است که سیگنال‌های کنترلی تولیدشده توسط عامل از نظر سخت‌افزاری قابل تحقق باشند و تنش بیش از حد به عملگر وارد نکنند. به همین منظور، در تابع پاداش چند مؤلفه جریمه در نظر گرفته شد که به بزرگی گشتاور، نرخ تغییر آن و همچنین تغییرات شتاب گشتاور مربوط می‌شوند. این اجزاء مانع از آن می‌شوند که عامل سیاست‌های کنترلی تهاجمی یا غیرواقعی بیاموزد.

– جریمه‌ی بزرگی گشتاور

اگرچه استفاده از گشتاورهای بزرگ می‌تواند از نظر تئوری منجر به بازگشت سریع به تعادل شود، اما از دیدگاه عملی به دلیل

^۱ Alignment Reward

محدودیت‌های عملگر مطلوب نیست. از این رو بزرگی گشتاور نرمال شده با رابطه ۲۶ جریمه می‌شود.

$$R_{\tau} = \left(\frac{\tau_{React}}{\tau_{max}} \right)^2 \quad 26$$

- جریمه‌ی نرخ تغییر گشتاور

برای جلوگیری از تغییرات ناگهانی در سیگنال کنترلی، نرخ تغییر گشتاور بین دو گام متوالی به صورت رابطه ۲۷ جریمه می‌گردد.

$$R_{\Delta\tau} = \left(\frac{\tau_{React} - \tau_{React_{prev}}}{\tau_{max} / \Delta t} \right)^2 \quad 27$$

- جریمه‌ی شتاب تغییر گشتاور

برای جلوگیری از نوسانات با فرکانس بالا در سیگنال کنترلی، تغییرات مرتبه دوم گشتاور نیز جریمه می‌شود. این جریمه مطابق رابطه ۲۸ تعریف می‌شود.

$$R_{\Delta^2\tau} = \left(\frac{\tau_{React} - 2\tau_{React_{prev}} + \tau_{React_{2prev}}}{\tau_{max} / (\Delta t^2)} \right)^2 \quad 28$$

- جریمه نزدیک شدن به مرز اشباع

برای جلوگیری از عملکرد عامل در شرایط نزدیک به حد توان عملگر (ناحیه اشباع)، مؤلفه‌ای از تابع پاداش به صورت جریمه تعریف شد.

$$R_{sat} = \left(\frac{|\tau_{React}|}{\tau_{max}} \right)^6 \quad 29$$

انتخاب توان ششم باعث می‌شود مقدار جریمه در محدوده ایمن بسیار کم، اما با نزدیک شدن به مرز اشباع به صورت نمایی افزایش یابد. بدین ترتیب، عامل از تولید گشتاورهای بیش از حد اجتناب کرده و عملکرد سیستم در ناحیه ایمن حفظ می‌شود. این طراحی با هدف افزایش ایمنی، کاهش استهلاک عملگر و جلوگیری از ناپایداری دینامیکی در نظر گرفته شده است. علاوه بر این تدبیر نرم افزاری در تابع پاداش، به منظور شبیه سازی دقیق محدودیت‌های سخت افزاری موتور، گشتاور اعمالی به چرخ عکس العمل در مدل فیزیکی شبیه ساز نیز به صورت اکید با یک بلوک اشباع (Saturation) در محدوده‌ی فیزیکی ± 1 نیوتن متر (یعنی $-1 \leq \tau_{React} \leq 1$) محدود شده است تا مانع از اعمال گشتاورهای غیرواقعی توسط کنترل کننده گردد.

- جریمه سرعت زاویه‌ای چرخ عکس العمل

یکی از محدودیت‌های مهم سیستم به سرعت زاویه‌ای چرخ عکس العمل (ω) مربوط می‌شود. تجاوز از حد مجاز آن علاوه بر خطر ناپایداری دینامیکی، موجب آسیب سخت افزاری و کاهش عمر عملگر می‌گردد. برای در نظر گرفتن این مسئله، دو سطح جریمه تعریف شد.

جریمه نرم^۱: زمانی که $|\omega|$ از مقدار ایمن ω_{soft} فراتر رود، اما هنوز به بیشینه مجاز ω_{max} نرسیده باشد، مقدار جریمه طبق رابطه زیر به صورت تدریجی افزایش می‌یابد.

$$R_{soft} = \begin{cases} 0 & |\omega| \leq \omega_{soft} \\ \left(\frac{|\omega_{React}| - \omega_{soft}}{\omega_{max} - \omega_{soft}} \right)^2 & \omega_{soft} < |\omega| < \omega_{max} \end{cases} \quad 30$$

جریمه شکست^۲: در صورت عبور سرعت زاویه‌ای از حد بیشینه، اپیزود خاتمه یافته و جریمه ثابت سنگینی مطابق رابطه زیر اعمال می‌شود.

$$R_{overspeed} = -5 \quad 31$$

- پاداش بقای اپیزود و شروط پایانی

برای افزایش پایداری عامل، در هر گام زمانی یک پاداش بقا کوچک در نظر گرفته شد تا عامل برای ادامه اپیزود و اجتناب از واژگونی ترغیب شود. در صورت عبور زاویه غلتش از حد مجاز، اپیزود خاتمه یافته و جریمه سنگینی اعمال می‌گردد:

$$R_{fall} = -15 \quad 32$$

اگر اپیزود بدون شکست به پایان برسد، عامل پاداش مثبت قابل توجهی دریافت می‌کند.

$$R_{success} = +15 \quad 33$$

این ساختار، عامل را به حفظ تعادل پایدار و ادامه حرکت ایمن در طول فرایند یادگیری تشویق می‌کند.

- تابع پاداش کلی

پاداش نهایی در هر گام زمانی به صورت ترکیب وزن دار مؤلفه‌های مختلف تابع پاداش محاسبه شد.

$$R_{total} = w_{good} R_{good} + w_{align} R_{align} + w_{lyap} R_{lyap} - w_{\tau} R_{\tau} - w_{\Delta\tau} R_{\Delta\tau} - w_{\Delta^2\tau} R_{\Delta^2\tau} - w_{sat} R_{sat} - w_{soft} R_{soft} + R_{alive} \quad 34$$

^۲ Overspeed Penalty

^۱ Soft Penalty

$R_{success}$	+120	Terminal bonus for successful navigation or early stabilization
$R_{overspeed}$	-80	Terminal penalty for exceeding maximum safe angular speed

Table 5. Components of equations 21-33

جدول ۵. مؤلفه‌ها و اجزای معادلات ۲۱ تا ۳۳

Symbol	Unit	Description
θ_{bike}	rad	Body roll angle
θ_{well}	rad	Desired region for roll angle
$\dot{\theta}_{bike}$	rad / s	Roll angular velocity
$\dot{\theta}_{well}$	rad / s	Desired region for roll angular velocity
R_{good}	dimensionless	Reward related to maintaining upright balance
$L(t)$	rad / s	Lyapunov function
R_{lyap}	dimensionless	Reward for Lyapunov energy reduction
τ_{React}	$N \cdot m$	Control input
$\tau_{React_{prev}}$	$N \cdot m$	Previous-step control input
$\tau_{React_{2_{prev}}}$	$N \cdot m$	Control input two steps earlier
Δt	s	Sampling time
τ_{max}	$N \cdot m$	Maximum allowable torque
R_{τ}	dimensionless	Penalty for torque magnitude
$R_{\Delta\tau}$	dimensionless	Penalty for torque-rate change
$R_{\Delta^2\tau}$	dimensionless	Penalty for torque second derivative
ω_{React}	rad / s	Reaction-wheel angular velocity
ω_{soft}	rad / s	Safe-speed threshold for reaction wheel
ω_{max}	rad / s	Maximum allowable reaction-wheel angular speed
R_{soft}	dimensionless	Penalty for nearing the soft speed limit
$R_{overspeed}$	dimensionless	Penalty for exceeding safe angular speed

۳.۲.۲. روش‌های یادگیری تقویتی

افزونه طراحی یادگیری تقویتی متلب^۲ مجموعه‌ای از الگوریتم‌ها را ارائه می‌دهد؛ از جمله Q-Learning و DQN برای فضای عمل گسسته، TD3، DDPG، PPO و SAC برای فضای عمل پیوسته. با توجه به پیوسته بودن فضای عمل در این پژوهش، الگوریتم گرادین

در هر گام، مقدار R_{total} به‌صورت وزن‌دار از مؤلفه‌های پاداش و جریمه محاسبه و در طول اپیزود تجمعی^۱ می‌شود. در پایان اپیزود، در صورت موفقیت عامل در حفظ تعادل، پاداش +15 و در صورت واژگونی، جریمه -15 به مجموع افزوده می‌گردد. این ساختار سیگنالی روشن از موفقیت یا شکست اپیزود ارائه داده و فرآیند یادگیری را به سمت رفتار پایدار هدایت می‌کند.

همان‌طور که پیش‌تر اشاره شد، طراحی دقیق و تنظیم متناسب مؤلفه‌های تابع پاداش، نقش حیاتی در همگرایی الگوریتم‌های یادگیری تقویتی دارد. در رابطه (۳۴)، ساختار کلی تابع پاداش به صورت ترکیبی خطی از پاداش‌ها و جریمه‌های مختلف ارائه گردید. به‌منظور شفاف‌سازی کامل فرآیند آموزش و فراهم‌سازی امکان بازتولید نتایج پژوهش، مقادیر عددی تمامی ضرایب وزنی و پاداش‌های پایانی به‌کاررفته در محیط شبیه‌سازی، در جدول ۴ استخراج و ارائه شده است. این مقادیر طی فرآیند آموزش و از طریق آزمون و خطا به گونه‌ای بهینه شده‌اند که تعادلی مناسب میان پایداری سریع، مصرف انرژی بهینه و نرمی سیگنال‌های کنترلی ایجاد نمایند.

Table 4. Numerical values of the weight coefficients and terminal rewards used in the reward function

جدول ۴. مقادیر عددی ضرایب وزنی و پاداش‌های پایانی استفاده‌شده در تابع پاداش

Symbol	Value	Description
W_{good}	2	Weight for maintaining upright balance
W_{align}	2.5	Weight for alignment and returning to upright state
W_{lyap}	1.6	Coefficient for Lyapunov energy reduction reward
W_{τ}	0.012	Base penalty coefficient for control input (torque) magnitude
$W_{\Delta\tau}$	0.03	Base penalty coefficient for torque rate-of-change
$W_{\Delta^2\tau}$	0.015	Base penalty coefficient for torque second derivative
W_{sat}	0.04	Penalty coefficient for nearing actuator saturation
W_{soft}	0.6	Penalty coefficient for exceeding the soft speed limit
R_{alive}	0.03	Survival bonus per second
R_{fall}	-150	Terminal penalty for falling (exceeding fail angle)

^۲ Reinforcement Learning Design

^۱ Cumulative Reward

$$x = \begin{bmatrix} \theta_{bike} \\ \dot{\theta}_{bike} \\ \omega_{React} \end{bmatrix} \quad 38$$

که در آن θ_{bike} زاویه‌ی غلتش، $\dot{\theta}_{bike}$ نرخ زاویه‌ای غلتش و ω_{React} سرعت زاویه‌ای چرخ عکس‌العمل است. ورودی کنترلی u گشتاور اعمال شده به چرخ عکس‌العمل است. ماتریس‌های فضای حالت عبارتند از:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ mgh/J_b & -b_\phi/J_b & 0 \\ 0 & 0 & -b_\omega/J_\omega \end{bmatrix} \quad 39$$

$$B = \begin{bmatrix} 0 \\ -1/J_b \\ 1/J_\omega \end{bmatrix} \quad 40$$

پارامترهای فیزیکی سیستم از طریق مدل سه‌بعدی طراحی شده در نرم‌افزار سالدورکس استخراج شده‌اند و مقادیر آن‌ها در جدول ۶ ارائه شده است. با جایگذاری این مقادیر، ماتریس‌های فضای حالت سیستم به صورت عددی محاسبه می‌گردند.

به منظور طراحی کنترل کننده، ماتریس‌های وزن دهی تابع هزینه (Q و R) با هدف تأکید بر پایداری سریع زاویه غلتش و جلوگیری از واگرایی سرعت زاویه‌ای چرخ عکس‌العمل تنظیم شدند. عناصر قطری ماتریس Q به ترتیب متناظر با کیفر زاویه غلتش (ϕ)، نرخ تغییر زاویه غلتش ($\dot{\phi}$) و سرعت زاویه‌ای چرخ عکس‌العمل (ω) هستند. برای سناریوی حرکت در مسیر مستقیم، وزن‌های زیر انتخاب شدند.

$$Q_1 = \text{diag}(1500, 3, 100), \quad R_1 = 3000$$

وزن بالای $Q_1(1,1) = 1500$ بیانگر اولویت اول پایداری زاویه غلتش و $R_1 = 3000$ محدودکننده‌ی گشتاور اعمال شده به چرخ عکس‌العمل است. پس از حل معادله جبری ریکاتی، ماتریس بهره کنترلی بهینه برای این سناریو به صورت زیر محاسبه گردید.

$$K_{LQR,1} = [-96.90 \quad -75.72 \quad 0.0881]$$

مقادیر ویژه سیستم حلقه بسته ($A_{cl} = A - BK_1$) بصورت زیر هستند.

$$\lambda_1 = -117.94, \quad \lambda_2 = -1.240, \quad \lambda_3 = -1.280$$

سیاست قطعی عمیق دوقلو با تأخیر و گرادیان سیاست قطعی عمیق به عنوان گزینه نهایی انتخاب شد.

۴.۲.۲. تنظیم ابرمولفه‌ها^۱

تنظیم ابرمولفه‌ها از مراحل کلیدی در فرایند آموزش عامل یادگیرنده است و تأثیر مستقیمی بر دقت و پایداری مدل دارد. در این پژوهش، مقادیر اولیه بر اساس مطالعات پیشین انتخاب و با استفاده از روش جستجوی شبکه‌ای و آزمایش‌های تجربی بهینه‌سازی شدند [26]. از مهم‌ترین پارامترها می‌توان به نرخ یادگیری، تعداد تکرارهای آموزشی، اندازه دسته داده و ساختار شبکه اشاره کرد. برای جلوگیری از بیش‌برازش^۲ نیز از راهبردهای اعتبارسنجی متقابل و توقف زودهنگام بهره گرفته شد.

۳.۲. کنترل کننده خطی درجه دوم بهینه

کنترل کننده خطی درجه دوم بهینه یک روش کنترل بازخورد حالت بهینه برای سیستم‌های خطی است که هدف آن کمینه‌سازی یک تابع هزینه درجه دوم است. این تابع هزینه هم انحراف متغیرهای حالت و هم تلاش کنترلی را جریمه می‌کند. برای یک سیستم خطی گسسته‌زمان به شکل:

$$x_{(t+1)} = Ax_t + Bu_t \quad 35$$

تابع هزینه به صورت زیر تعریف می‌شود.

$$J = \sum_{t=0}^{N-1} (x_t^T Q x_t + u_t^T R u_t) + x_N^T Q_f x_N \quad 36$$

که در آن Q ، R و Q_f ماتریس‌های وزن دهی هستند (Q و Q_f نیمه‌معین مثبت و R معین مثبت). حل این مسئله بهینه‌سازی به قانون کنترل بازخورد خطی زیر منجر می‌شود.

$$u_t^* = -K_{(LQR)} x_t \quad 37$$

بهره‌ی K_{LQR} از حل معادلات ریکاتی (به صورت بازگشتی برای افق محدود یا از معادله جبری ریکاتی برای افق نامحدود) به دست می‌آید.

برای طراحی کنترل کننده‌ی LQR، دینامیک غلتش دوچرخه‌ی مجهز به چرخ عکس‌العمل در فرم فضای حالت بیان می‌شود. بردار حالت به صورت زیر تعریف می‌گردد.

^۱ Overfitting

^۲ Hyperparameters

جدول ۷. پارامترهای طراحی و مشخصات پایداری کنترل کننده LQR برای مسیرهای مستقیم و منحنی

Gain	Scenario	
	Straight Path	Curved Path
Q	diag(1500,3,100)	diag(15000,3,1000)
R	3000	3000
K_ϕ	-96.90	-304.11
$K_{\phi\dot{}}$	-75.72	-237.64
K_ω	0.0881	0.2846
λ_1	-117.94	-372.78
λ_2	-1.240	-1.241
λ_3	-1.280	-1.279

شایان ذکر است که مدل فضای حالت ارائه شده یک مدل ساده شده‌ی جاداشده برای دینامیک غلتش است که کوپلینگ صریح بین زاویه‌ی فرمان، سرعت پیشروی و غلتش را در بر نمی‌گیرد. این ساده‌سازی آگاهانه به منظور امکان‌پذیر ساختن طراحی تحلیلی کنترل کننده‌ی LQR انجام شده و در ادبیات کنترل وسایل دوچرخه‌ی متداول است [21], [3]. با این حال، اعتبارسنجی و ارزیابی کامل سیستم در محیط شبیه‌سازی چندجسمی Simscape انجام شده که تمام دینامیک غیرخطی کوپل شده، اثرات ژيروسکوپی و نیروهای گریز از مرکز را به‌طور کامل در خود دارد. این رویکرد، طراحی کنترل کننده بر پایه‌ی مدل ساده و آزمون در محیط دینامیکی کامل، تفاوت عملکردی LQR و الگوریتم‌های یادگیری تقویتی (که مستقیماً از محیط یاد می‌گیرند) را برجسته می‌سازد و یکی از نوآوری‌های اصلی چارچوب پیشنهادی را تشکیل می‌دهد.

۳. نتایج و بحث

در این بخش، عملکرد چارچوب کنترلی پیشنهادی در محیط شبیه‌سازی سیمولینک متلب ارزیابی می‌شود. به منظور ایجاد یک معیار مقایسه‌ی پایه، ابتدا کنترل کننده‌ی کلاسیک LQR که در بخش ۳.۲ طراحی گردید، در هر سناریو اجرا می‌شود. سپس عملکرد الگوریتم‌های هوشمند DDPG و TD3 با این معیار مقایسه می‌گردد. این ارزیابی در سه سناریو انجام می‌شود.

۱. حفظ تعادل در مسیر مستقیم

۲. ناوبری در حضور موانع (مسیر منحنی)

که بخش حقیقی منفی تمام آن‌ها پایداری مجانبی سیستم را در این سناریو تضمین می‌کند. برای سناریوی ناوبری در حضور موانع (مسیر منحنی)، به دلیل اعمال اغتشاشات ناشی از تغییر فرمان و نیاز به غلبه بر نیروهای جانبی هنگام دور زدن، پارامترهای طراحی متناسب با این شرایط بازتنظیم شدند:

$$Q_2 = \text{diag}(15000, 3, 1000), \quad R_2 = 3000$$

افزایش ده‌برابری وزن زاویه غلتش (از 1500 به 15000) و وزن سرعت چرخ عکس‌العمل (از 100 به 1000) موجب می‌شود کنترل کننده واکنش سریع‌تر و قوی‌تری نسبت به انحرافات جانبی داشته باشد. پس از حل معادله جبری ریکاتی، ماتریس بهره به‌صورت زیر محاسبه گردید.

$$K_{LQR,2} = [-304.11 \quad -237.64 \quad 0.2846]$$

مقادیر ویژه سیستم حلقه‌بسته بصورت زیر هستند که پایداری مجانبی سیستم در سناریوی ناوبری را تأیید می‌کند.

$$\lambda_1 = -372.78, \quad \lambda_2 = -1.241, \quad \lambda_3 = -1.279$$

همان‌طور که در جدول ۷ مشاهده می‌شود، در سناریوی ناوبری به دلیل نیاز به واکنش‌های کنترلی سریع‌تر و قوی‌تر چرخ عکس‌العمل برای دفع اغتشاشات، مقادیر بهره کنترلی نسبت به مسیر مستقیم افزایش قابل توجهی یافته‌اند.

Table 6. Physical parameters of the self-stabilizing bicycle system

جدول ۶. پارامترهای فیزیکی سیستم دوچرخه خودپایدارساز

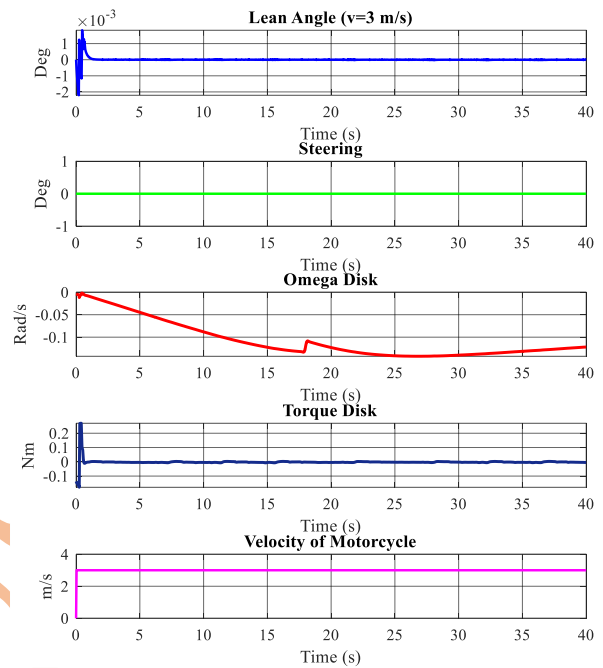
Description	Value	Unit
Bicycle assembly mass (m)	6.398	kg
Roll moment of inertia about Y-axis (J_a)	0.388	$\text{kg} \cdot \text{m}^2$
Reaction-wheel moment of inertia (J_w)	1.549×10^{-3}	$\text{kg} \cdot \text{m}^2$
Height of the center of mass (h)	0.104	m
Roll damping coefficient (b_ϕ)	0.05	$\text{N} \cdot \text{m} \cdot \text{s} / \text{rad}$
Reaction-wheel damping coefficient (b_w)	0.00573	$\text{N} \cdot \text{m} / \text{rad}$
Gravity magnitude (g)	9.81	m / s^2

Table 7. Design parameters and stability characteristics of the LQR controller for straight and curved paths

۳. مقاومت در برابر اغتشاش ناشی از دست‌انداز

۱.۳. حفظ تعادل بر روی خط مستقیم

در گام نخست، کنترل‌کننده LQR با بهره‌ی بر $K_{LQR} = [-96.8985 \quad -75.7215 \quad 0.0881]$ روی سیستم اعمال گردید. نتایج شبیه‌سازی در شکل ۹ نشان داده شده است.



شکل ۹. نتایج شبیه‌سازی سیستم با کنترل‌کننده LQR در سناریوی فرمان ثابت. نمودارها (از بالا به پایین) نشان‌دهنده: زاویه غلتش، ورودی فرمان، سرعت زاویه‌ای چرخ عکس‌العمل، گشتاور چرخ عکس‌العمل و سرعت رو به جلو هستند

Figure 9. Simulation results of the system employing the LQR controller under the fixed-steering scenario. The subplots illustrate (from top to bottom): lean angle, steering input, reaction-wheel angular velocity, reaction-wheel torque, and forward velocity

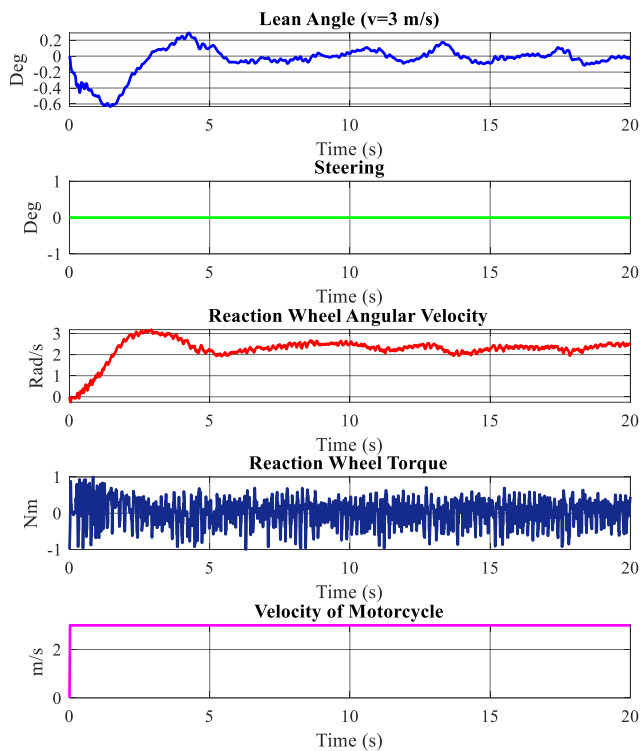
نتایج شبیه‌سازی کنترل‌گر LQR برای حفظ تعادل موتورسیکلت دوجرخ در حرکت مستقیم با سرعت ۳ متر بر ثانیه در شکل ۹ ارائه شده است. شایان ذکر است که انتخاب سرعت ۳ متر بر ثانیه با هدف بررسی پایداری در ناحیه بحرانی سرعت‌پایین (زیر ۵ متر بر ثانیه) صورت گرفته است، زیرا در این سرعت اثرات پایدارساز طبیعی هندسه وسیله ناچیز بوده و حفظ تعادل بیشترین چالش کنترلی را به همراه دارد. همان‌طور که مشاهده می‌شود، زاویه‌ی انحراف جانبی موتورسیکلت در پاسخ به اغتشاش اولیه در حدود 10^{-3} درجه، توسط کنترل‌کننده به سرعت میرا شده و در زمان کوتاهی به مقدار صفر همگرا می‌شود؛ این امر نشان‌دهنده‌ی پاسخ دینامیکی مطلوب و

پایداری مجانبی سیستم تحت کنترل LQR است. زاویه‌ی فرمان در طول کل شبیه‌سازی تقریباً صفر باقی می‌ماند، که بیانگر آن است که در حرکت مستقیم، مکانیزم تعادل عمدتاً از طریق گشتاور ژيروسکوپ دیسک چرخان و نه از طریق چرخش فرمان تأمین می‌گردد. سرعت زاویه‌ای دیسک روندی کاهشی تدریجی دارد که نشان‌دهنده‌ی عملکرد پیوسته‌ی دیسک برای ایجاد اثر ژيروسکوپ لازم جهت حفظ تعادل است. گشتاور اعمال‌شده به دیسک در محدوده‌ی $\pm 0.2 \text{ N}\cdot\text{m}$ محدود مانده و پس از خاموش شدن اغتشاش اولیه، به مقادیر بسیار کوچکی در اطراف صفر کاهش می‌یابد؛ این رفتار بیانگر مصرف انرژی پایین عملگر در حالت پایدار و تأیید ماهیت بهینه‌ی کنترل‌گر LQR از منظر تعادل بین عملکرد و هزینه‌ی کنترل است. در نهایت، سرعت خطی موتورسیکلت در تمام بازه‌ی شبیه‌سازی ثابت و برابر ۳ متر بر ثانیه حفظ شده است، که حاکی از جداسازی مناسب حلقه‌های کنترل سرعت و تعادل می‌باشد. به‌طور کلی، نتایج حاکی از آن است که کنترل‌گر LQR طراحی‌شده قادر است با اعمال ورودی‌های کنترلی محدود، پایداری دینامیکی موتورسیکلت را در حرکت مستقیم به‌صورت مؤثر و بهینه تضمین نماید.

در این بخش، دو الگوریتم DDPG و TD3 برای آموزش مدل کنترلی به‌منظور حفظ تعادل دوجرخه بر روی مسیر مستقیم به‌کار گرفته شدند. با تنظیم مناسب آبرمولفه‌ها، الگوریتم DDPG توانست در تعداد اپیزود کمتر به همگرایی برسد و نوسانات اولیه را به‌طور مؤثری کاهش دهد. در مقابل، الگوریتم TD3 اگرچه به اپیزودهای بیشتری برای همگرایی نیاز داشت، اما در نهایت رفتار پایدارتر و مقاوم‌تری در برابر نوسانات و اغتشاشات از خود نشان داد. نتایج شبیه‌سازی بیانگر آن است که DDPG از نظر سرعت یادگیری و TD3 از نظر پایداری نهایی عملکرد بهتری دارند.

به منظور اعتبارسنجی فرآیند آموزش و اطمینان از پایداری یادگیری، روند همگرایی الگوریتم‌ها مورد ارزیابی قرار گرفت. به‌عنوان نمونه، شکل ۱۰ منحنی پاداش تجمیعی را برای الگوریتم TD3 در طول بیش از ۷۰۰۰ اپیزود آموزشی نشان می‌دهد. همان‌طور که در روند صعودی خط میانگین متحرک (خط نارنجی‌رنگ) کاملاً مشهود است، عامل یادگیرنده بدون مواجهه با مشکل فراموشی فاجعه‌بار و با یک شیب یادگیری پیوسته، به سیاست کنترلی بهینه همگرا شده است.

Soft-update coefficient	0.05	0.05
Target standard deviation (initial)	0.2	0.2
Target standard deviation	0	0
Clipping bound	1×10^{-5}	1×10^{-5}
Standard-deviation decay rate	0.02	0.05
Actor learning rate	0.0001	0.0001
Critic learning rate	0.0001	0.0003
Successful episodes	6900	1250

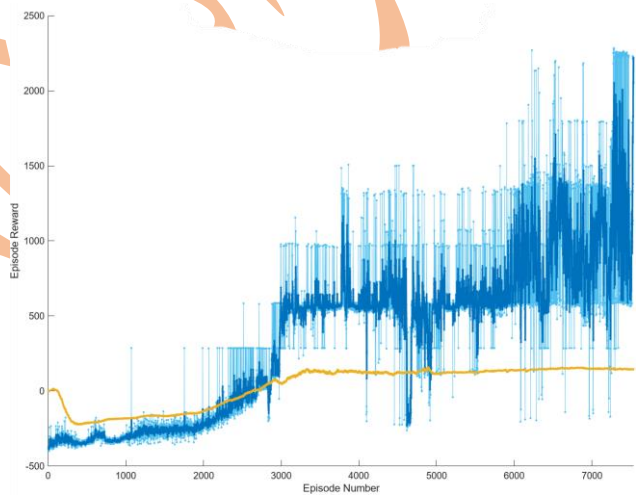


شکل ۱۱. نتایج شبیه‌سازی سیستم با الگوریتم یادگیری تقویتی DDPG در سناریوی فرمان ثابت، شامل زاویه غلتش، ورودی فرمان، سرعت رو به جلو، گشتاور چرخ عکس‌العمل و سرعت زاویه‌ای چرخ عکس‌العمل

Figure 11. Simulation results of the system employing reinforcement learning with the DDPG algorithm under the fixed-steering scenario, including roll angle, steering input, forward velocity, reaction-wheel torque, and reaction-wheel angular velocity

نتایج شبیه‌سازی (شکل ۱۱) نتایج شبیه‌سازی دوچرخه با زاویه فرمان ثابت برابر صفر تحت الگوریتم DDPG نشان می‌دهد که زاویه انحراف بدنه ابتدا دارای نوسانات گذرا است، اما به تدریج میرا شده و حول صفر پایدار می‌گردد. سرعت زاویه‌ای چرخ عکس‌العمل پس از افزایش اولیه در محدوده‌ای مشخص تثبیت شده و گشتاور کنترلی نیز در حدود ± 1 نیوتن‌متر باقی می‌ماند که عملکرد مناسب در حفظ

این روند پایدار آموزش، قابلیت بازتولید الگوریتم را در تعامل با محیط پیچیده‌ی چندجسمی سیمولینک تضمین می‌نماید. شایان ذکر است با توجه به اینکه محیط آموزش در این پژوهش یک شبیه‌سازی کامل چندجسمی سه‌بعدی در Simscape Multibody است و حل عددی معادلات کوپل‌شده غیرخطی هزینه محاسباتی بسیار بالایی دارد (هر دوره آموزش ۷۰۰۰ اپیزودی نیازمند چند روز پردازش مداوم پردازنده است)، ارزیابی آماری و تکرار چندباره آموزش با بذره‌ای تصادفی^۱ متعدد عملاً با محدودیت منابع محاسباتی مواجه بوده است؛ با این حال، شیب یادگیری پیوسته و عدم واگرایی در طول اپیزودها، صحت و تکرارپذیری رفتار عامل را به اثبات می‌رساند. ابرپارامترهای شبکه‌های یادگیری تقویتی برای حالت حرکت در مسیر مستقیم در جدول ۸ ارائه شده‌اند.



شکل ۱۰. منحنی همگرایی پاداش تجمعی اپیزودها در طول فرآیند آموزش عامل TD3

Figure 10. Convergence curve of the cumulative episode reward during the TD3 agent training process

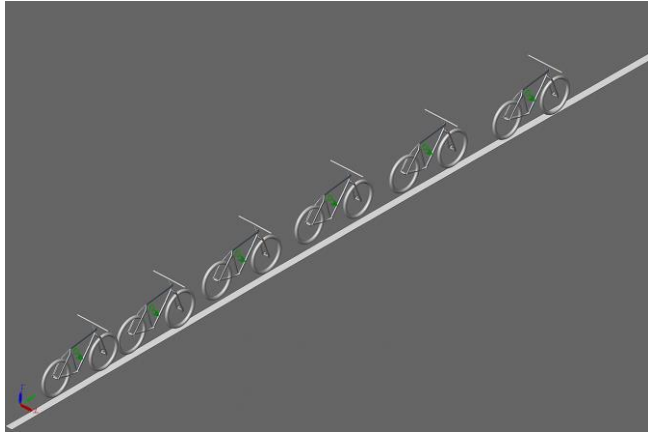
Table 8. Hyperparameters of the reinforcement learning framework for maintaining balance along a straight path

جدول ۸. ابرپارامترهای شبکه‌ی یادگیری تقویتی برای حفظ تعادل در مسیر مستقیم

Parameter	TD3	DDPG
Sampling time	0.02	0.02
Discount factor	0.995	0.995
Replay buffer size	2×10^6	2×10^6
Actor threshold	1	1
Critic threshold	1	1

^۱ Random Seeds

مقایسه با الگوریتم DDPG نشان می‌دهد که TD3 از نظر پایداری نهایی، خطای ماندگار و زمان نشست عملکرد بهتری دارد. اضافه‌جهش اولیه کمتر و کنترل گشتاور چرخ عکس‌العمل با پایداری و نرمی بیشتری انجام می‌شود. بنابراین، TD3 نسبت به DDPG قابلیت بالاتری در بهبود پایداری دینامیکی دوچرخه تحت شرایط زاویه فرمان ثابت دارد و گزینه مناسب‌تری برای طراحی کنترل‌کننده محسوب می‌شود.

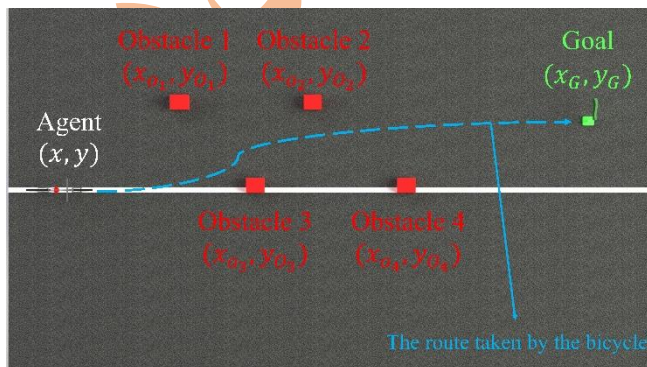


شکل ۱۳. مسیر حرکت در حین حفظ تعادل در یک مسیر مستقیم

Figure 13. The trajectory during balance maintenance along a straight path

۲.۳. حفظ تعادل بر روی مسیر منحنی

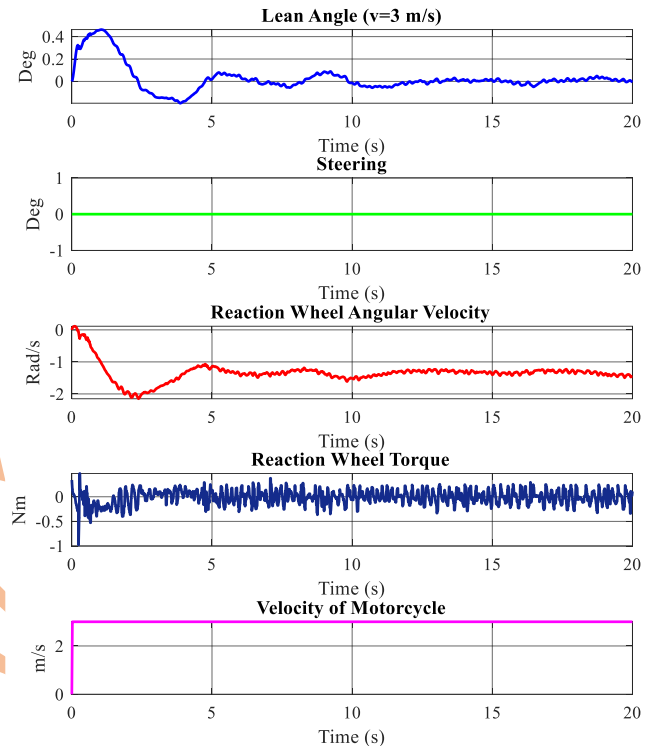
الگوریتم نوبری در حضور چهار مانع نیز تحت همان شرایط آزمایشی پیشین اجرا شد و آموزش مدل با استفاده از دو الگوریتم یادگیری تقویتی DDPG و TD3 انجام گرفت. در این بخش، همان چارچوب شبیه‌سازی و پارامترهای کنترلی مراحل قبلی به کار گرفته شد تا امکان مقایسه عملکرد سیستم در شرایط وجود چند مانع فراهم شود.



شکل ۱۴. دوچرخه خودپایدارساز در حضور چهار مانع

Figure 14. The self-stabilizing bicycle in the presence of four obstacles

پایداری دینامیکی را تأیید می‌کند. سرعت زاویه‌ای چرخ عقب ثابت مانده و حرکت یکنواخت سیستم را نشان می‌دهد. این نتایج بیانگر آن است که حتی با ثابت بودن زاویه فرمان، اعمال گشتاور کنترلی چرخ عکس‌العمل به بهبود پایداری حرکتی دوچرخه کمک می‌کند.



شکل ۱۲. نتایج شبیه‌سازی سیستم با الگوریتم یادگیری تقویتی TD3 در سناریوی فرمان ثابت، شامل زاویه غلتش، ورودی فرمان، سرعت رو به جلو، گشتاور چرخ عکس‌العمل و سرعت زاویه‌ای چرخ عکس‌العمل

Figure 12. Simulation results of the system employing reinforcement learning with the TD3 algorithm under the fixed-steering scenario, including roll angle, steering input, forward velocity, reaction-wheel torque, and reaction-wheel angular velocity

در شکل ۱۲، نتایج شبیه‌سازی دوچرخه در شرایطی که زاویه فرمان ثابت و برابر صفر است، برای الگوریتم TD3 ارائه شده است. مطابق نمودارها، زاویه انحراف بدنه ابتدا به بیشینه حدود 0.45 درجه می‌رسد و سپس با نرخ میرایی مناسب کاهش یافته و به حوالی صفر می‌رسد، جایی که با نوسانات بسیار کوچک پایدار می‌ماند. سرعت زاویه‌ای چرخ عکس‌العمل پس از کاهش اولیه تثبیت شده و گشتاور کنترلی در بازه‌ای محدود بدون اشباع باقی می‌ماند. سرعت زاویه‌ای چرخ عقب نیز ثابت مانده و حرکت یکنواخت دوچرخه را نشان می‌دهد.

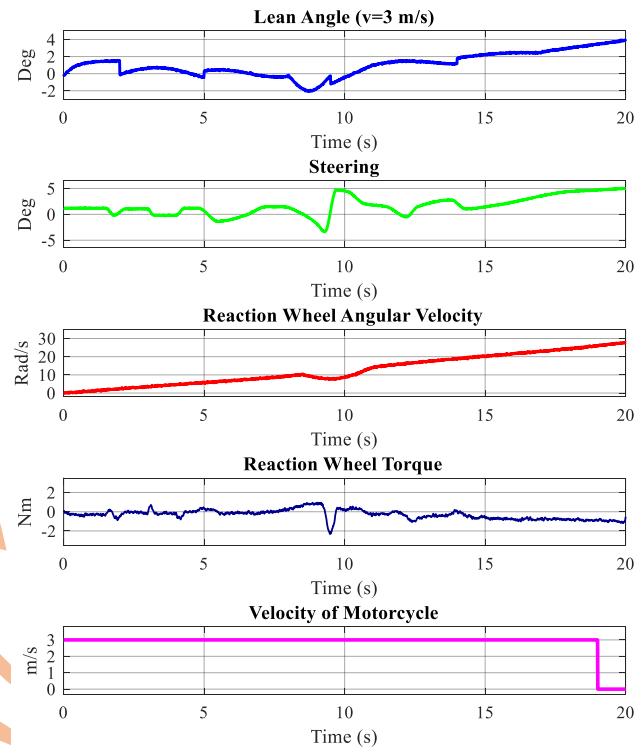
از مسیر منحنی است؛ به‌ویژه در حدود $t \approx 9-10s$ ثانیه که یک پیک قابل توجه در هر دو سیگنال زاویه‌ی انحراف و فرمان مشاهده می‌شود و بیانگر عبور از بخشی با انحنای تندتر در مسیر است. سرعت زاویه‌ای چرخ عکس‌العمل در طول کل شبیه‌سازی به‌طور پیوسته از صفر به حدود 30 rad/s افزایش می‌یابد؛ این رفتار نشان می‌دهد که برای تأمین گشتاور ژيروسکوپي لازم جهت مقابله با لحظه‌های واژگونی ناشی از انحنای مسیر، چرخ عکس‌العمل به‌تدریج شتاب می‌گیرد. گشتاور اعمال شده به چرخ عکس‌العمل عمدتاً در محدوده‌ی $\pm 2 \text{ N}\cdot\text{m}$ باقی می‌ماند، هرچند در لحظه‌ی عبور از قوس تند ($t \approx 10 \text{ s}$) یک پیک منفی به مقدار تقریبی $-2.5 \text{ N}\cdot\text{m}$ مشاهده می‌شود که نشان‌دهنده‌ی افزایش تقاضای کنترلی در آن بازه است. سرعت خطی موتورسیکلت در تمام مدت شبیه‌سازی ثابت و برابر 3 m/s حفظ شده و در انتهای شبیه‌سازی ($t \approx 20 \text{ s}$) به صفر می‌رسد که مرتبط با پایان مسیر تعریف شده است. به‌طور کلی، نتایج نشان می‌دهد که کنترل گر LQR قادر است تعادل موتورسیکلت را در مسیر منحنی نیز حفظ نماید، اما در مقایسه با حالت حرکت مستقیم، فعالیت بیشتری از عملگرها (فرمان و چرخ عکس‌العمل) مورد نیاز است و اثر تجمعی سرعت زاویه‌ای چرخ عکس‌العمل (اشباع احتمالی در بلندمدت) به‌عنوان یک چالش مطرح می‌باشد. ابرپارامترهای شبکه‌های یادگیری تقویتی برای سناریوی حرکت در مسیر منحنی در جدول ۹ ارائه شده‌اند.

Table 9. Hyperparameters of the reinforcement learning framework for maintaining balance along a curved path

جدول ۹. ابرپارامترهای شبکه‌ی یادگیری تقویتی برای حفظ تعادل در مسیر منحنی

Parameter	DDPG	TD3
Sampling time	0.02	0.02
Discount factor	1	0.99
Replay buffer size	2×10^6	1×10^6
Gradient threshold (actor)	∞	5
Gradient threshold (critic)	1	5
Target soft-update coefficient	0.05	0.05
Initial standard deviation	0.2	0.3
Minimum standard deviation	0	0
Standard-deviation decay rate	1×10^{-5}	5×10^{-5}
Learning rate (actor)	0.0001	0.0001
Learning rate (critic)	0.0003	0.0001

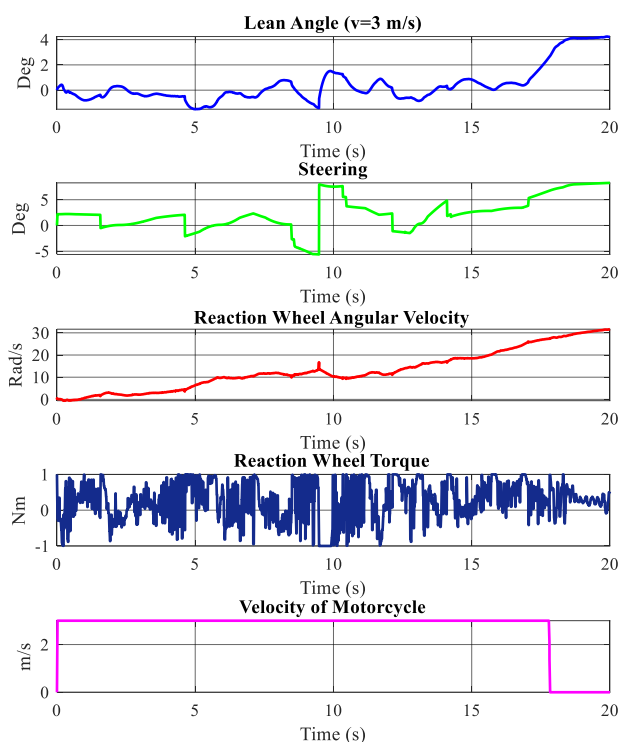
به منظور مقایسه، ابتدا کنترل‌کننده‌ی LQR با بهره‌ی $K_{LQR} = [-304.1056 \quad -237.6437 \quad 4.2846]$ در این سناریو اجرا گردید. نتایج در شکل ۱۵ ارائه شده است.



شکل ۱۵. نتایج شبیه‌سازی سیستم با کنترل‌کننده LQR در سناریوی کنترل ناوبری (بدون اغتشاش). نمودارها (از بالا به پایین) نشان‌دهنده: زاویه غلتش، ورودی فرمان، سرعت زاویه‌ای چرخ عکس‌العمل و گشتاور چرخ عکس‌العمل هستند

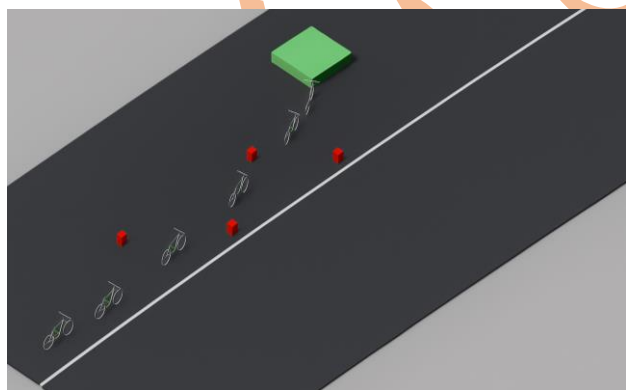
Figure 15. Simulation results of the system employing the LQR controller under the navigation-control scenario (without disturbance). The subplots illustrate (from top to bottom): lean angle, steering input, reaction-wheel angular velocity, reaction-wheel torque

نتایج شبیه‌سازی کنترل گر LQR برای حفظ تعادل موتورسیکلت در حرکت روی مسیر منحنی با سرعت ۳ متر بر ثانیه در شکل ۱۵ ارائه شده است. برخلاف حالت حرکت مستقیم، در این سناریو زاویه‌ی انحراف جانبی موتورسیکلت در طول شبیه‌سازی غیرصفر بوده و در بازه‌ی تقریبی 4° تا -2° نوسان می‌کند. این رفتار کاملاً مورد انتظار است، چراکه در حرکت روی مسیر منحنی، موتورسیکلت باید به‌سمت داخل قوس کج شود تا نیروی گریز از مرکز با مؤلفه‌ی افقی وزن متعادل گردد و تعادل دینامیکی برقرار شود. زاویه‌ی فرمان نیز برخلاف حالت مستقیم، فعالانه تغییر می‌کند و در محدوده‌ی $\pm 5^\circ$ نوسان دارد که نشانه‌ی دخالت مستقیم مکانیزم فرمان برای پیروی



شکل ۱۷. نتایج شبیه‌سازی سیستم با الگوریتم یادگیری تقویتی TD3 در سناریوی کنترل ناوبری (بدون اغتشاش). نمودارها (از بالا به پایین) نشان‌دهنده: زاویه غلتش، ورودی فرمان، سرعت زاویه‌ای چرخ عکس‌العمل، گشتاور چرخ عکس‌العمل و سرعت رو به جلو هستند.

Figure 17. Simulation results of the system employing reinforcement learning with the TD3 algorithm under the navigation-control scenario (without disturbance). The subplots illustrate (from top to bottom): lean angle, steering input, reaction-wheel angular velocity, reaction-wheel torque and forward velocity

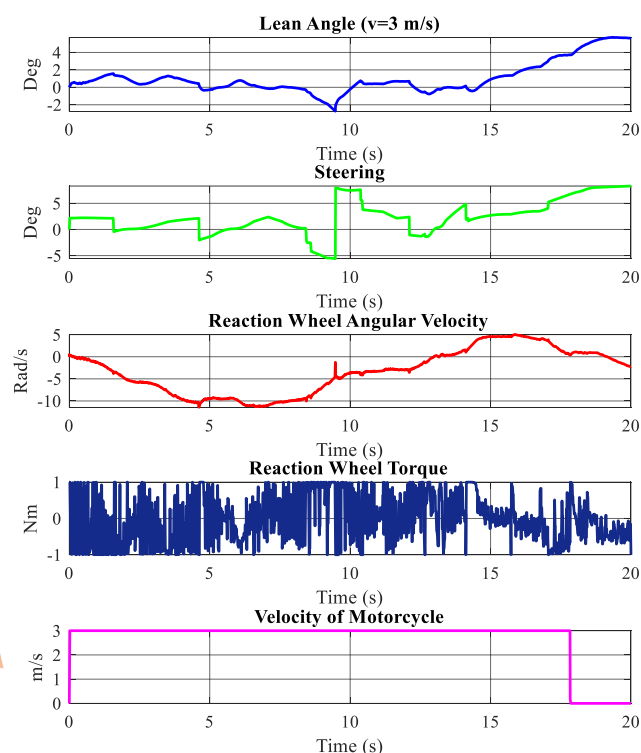


شکل ۱۸. محیط شبیه‌سازی الگوریتم ناوبری، شامل مسیر حرکت، موانع (به رنگ قرمز) و هدف (به رنگ سبز) برای دوچرخه خودپایدارساز

Figure 18. The navigation-algorithm simulation environment, including the trajectory, obstacles (in red), and the target destination (in green) for the self-stabilizing bicycle

مطابق شکل ۱۸، هر دو الگوریتم DDPG و TD3 توانستند در سرعت خطی ثابت ۳ متر بر ثانیه تعادل دوچرخه را حفظ کنند.

Successful episodes	3579	7537
---------------------	------	------



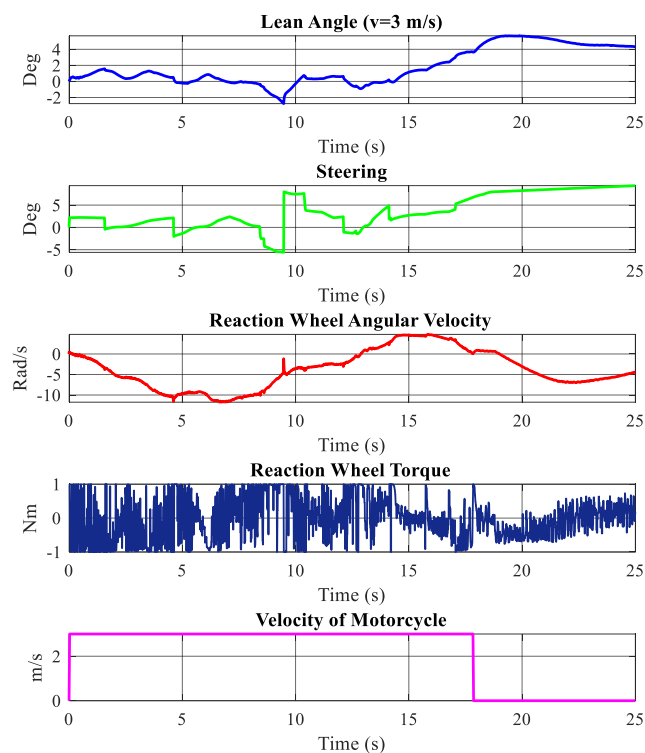
شکل ۱۶. نتایج شبیه‌سازی سیستم با الگوریتم یادگیری تقویتی DDPG در سناریوی کنترل ناوبری (بدون اغتشاش). نمودارها (از بالا به پایین) نشان‌دهنده: زاویه غلتش، ورودی فرمان، سرعت زاویه‌ای چرخ عکس‌العمل، گشتاور چرخ عکس‌العمل و سرعت رو به جلو هستند

Figure 16. Simulation results of the system employing reinforcement learning with the DDPG algorithm under the navigation-control scenario (without disturbance). The subplots illustrate (from top to bottom): lean angle, steering input, reaction-wheel angular velocity, reaction-wheel torque and forward velocity

امر برتری محسوس رویکرد یادگیری تقویتی را در مواجهه با دینامیک غیرخطی و مانورهای حرکتی اثبات می‌نماید

۳.۳. ارزیابی مقاومت سیستم در برابر اغتشاش

به منظور ارزیابی مقاومت^۱ سیستم‌های کنترلی در برابر اغتشاشات محیطی، یک دست‌انداز به عنوان نویز خارجی در مسیر حرکت دوچرخه اعمال شد. در این سناریو، الگوریتم‌های DDPG (شکل ۱۹) و TD3 (شکل ۲۰) بررسی می‌شوند.



شکل ۱۹. نتایج شبیه‌سازی سیستم با الگوریتم یادگیری تقویتی DDPG در سناریوی کنترل ناوبری با اغتشاش فیزیکی دست‌انداز. نمودارها (از بالا به پایین) نشان‌دهنده: زاویه غلتش، ورودی فرمان، سرعت زاویه‌ای چرخ عکس‌العمل، گشتاور چرخ عکس‌العمل و سرعت رو به جلو هستند

Figure 19. Simulation results of the system employing reinforcement learning with the DDPG algorithm under the navigation-control scenario with a speed bump disturbance. The subplots illustrate (from top to bottom): lean angle, steering input, reaction-wheel angular velocity, reaction-wheel torque and forward velocity

همان‌گونه که در شکل ۱۶ نشان داده شده، زاویه انحراف حول صفر با نوسانات محدود پایدار مانده، سرعت زاویه‌ای چرخ عکس‌العمل در بازه تقریباً 5 تا -10 رادیان بر ثانیه و گشتاور کنترلی در حدود ± 1 نیوتن‌متر باقی مانده است. در شکل ۱۷، پس از تغییر زاویه فرمان برای دور زدن مانع در زمان 9.5 ثانیه، بازبازی سریع و نرم مشاهده می‌شود، اگرچه سرعت چرخ عکس‌العمل تا حدود 30 رادیان بر ثانیه افزایش یافته است. در سناریوی مسیر منحنی با موانع، هر دو الگوریتم DDPG و TD3 موفق شدند دوچرخه را به محدوده‌ی هدف تعیین‌شده برسانند و از تمامی موانع اجتناب کنند؛ بدین ترتیب نرخ موفقیت ناوبری در سناریوهای آزمون‌شده به ۱۰۰٪ رسید.

در فرایند یادگیری، DDPG پس از 3579 اپیزود و TD3 پس از 7537 اپیزود به سیاست پایدار رسیدند. خطای ریشه‌میانگین مربع زاویه انحراف در بازه 0 تا 18 ثانیه 1.7 درجه برای DDPG و 1.23 درجه برای TD3 به‌دست آمد. بنابراین، TD3 در دفع اغتشاش و کاهش رانش انتهای عملکرد بهتری دارد، اما با مصرف انرژی بالاتر و سرعت بیشتر چرخ عکس‌العمل همراه است. DDPG با تعداد اپیزود کمتر و سرعت زاویه‌ای محدودتر برای کاربردهای دارای محدودیت سخت‌افزاری مناسب‌تر است.

علاوه بر مقایسه‌ی دو الگوریتم یادگیری، ارزیابی عملکرد سیستم با کنترلر کلاسیک LQR در دو سناریوی حرکت مستقیم (شکل ۹) و مسیر منحنی (شکل ۱۵) تفاوت‌های چشمگیری را آشکار می‌سازد. اگرچه کنترلر LQR بر اساس مدل خطی‌شده قادر به حفظ تعادل سیستم در مسیر مستقیم است، اما با شروع مانورهای ناوبری و اعمال پیوسته‌ی زاویه‌ی فرمان در مسیر منحنی، به دلیل ماهیت به‌شدت غیرخطی دینامیک دوچرخه، نوسانات بالایی در زاویه‌ی غلتش رخ داده و گشتاور چرخ عکس‌العمل رفتار پرنوسانی از خود نشان می‌دهد. در مقابل، کنترلرهای (یادگیری تقویتی عمیق) الگوریتم‌های (DDPG و TD3) بدون نیاز به خطی‌سازی مدل، تطبیق‌پذیری بسیار بالاتری از خود نشان داده و با اعمال گشتاورهای نرم‌تر، زاویه‌ی غلتش را با خطای کمتری حول نقطه‌ی تعادل حفظ می‌کنند. این

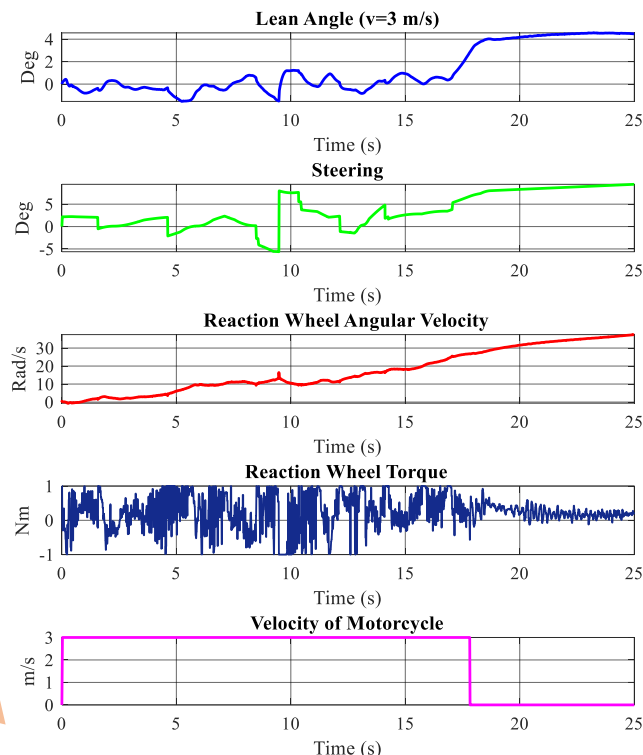
^۱ Robustness

بنابراین، در سناریوی ناوبری همراه با اغتشاش، الگوریتم DDPG به دلیل مدیریت بهینه‌تر سرعت چرخ عکس‌العمل و جلوگیری از اشباع آن، پایداری عملیاتی بلندمدت بالاتری ارائه می‌دهد و برای پیاده‌سازی سخت‌افزاری مناسب‌تر است.

۳.۳. ارزیابی مقاومت در برابر عدم قطعیت پارامتری و نویز اندازه‌گیری

به منظور پاسخ‌گویی به محدودیت‌های عملی و ارزیابی مقاومت چارچوب کنترلی پیشنهادی در شرایط واقعی، یک سناریوی تست مقاومت طراحی و پیاده‌سازی گردید. برای این منظور، به جای یک مسیر ناوبری نقطه‌به‌نقطه، یک تابع سینوسی به ورودی فرمان اعمال گردید. دلیل این انتخاب آن است که فرمان سینوسی به عنوان یک اغتشاش دینامیکی مداوم (مشابه مانورهای مارپیچ) عمل می‌کند و سیستم را به طور پیوسته در حالت گذار نگه می‌دارد؛ این امر امکان بررسی خالص پایداری غلتش را مستقل از منطق توقف الگوریتم ناوبری فراهم می‌سازد. در این سناریوی پویا، دو آزمون مقاومت مجزا انجام شد.

ابتدا، برای مدل‌سازی عدم قطعیت جرمی (نظیر اضافه شدن بار یا تغییر در سخت‌افزار)، جرم بدنه دوچرخه از 0.444 کیلوگرم به 1.5 کیلوگرم افزایش یافت که معادل یک عدم قطعیت جرمی ۱۶ درصدی در وزن کل سیستم است. همان‌طور که در شکل ۲۱ مشاهده می‌شود، الگوریتم TD3 بدون نیاز به بازتنظیم پارامترها، به خوبی توانست این عدم قطعیت را جذب کرده و با وجود مانور سینوسی، خطای غلتش را در محدوده بسیار امن ± 1.5 درجه حفظ کند.

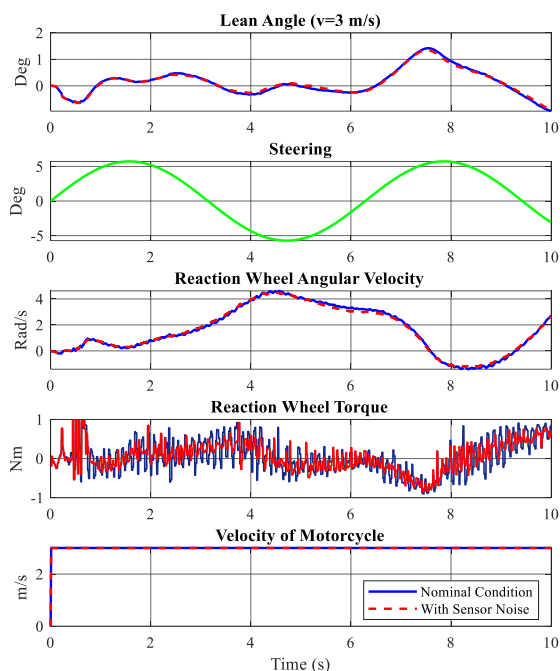


شکل ۲۰. نتایج شبیه‌سازی سیستم با الگوریتم یادگیری تقویتی TD3 در سناریوی کنترل ناوبری با اغتشاش فیزیکی دست‌انداز. نمودارها (از بالا به پایین) نشان‌دهنده: زاویه غلتش، ورودی فرمان، سرعت زاویه‌ای چرخ عکس‌العمل، گشتاور چرخ عکس‌العمل و سرعت رو به جلو هستند.

Figure 20. Simulation results of the system employing reinforcement learning with the TD3 algorithm under the navigation-control scenario with a speed bump disturbance. The subplots illustrate (from top to bottom): lean angle, steering input, reaction-wheel angular velocity, reaction-wheel torque and forward velocity

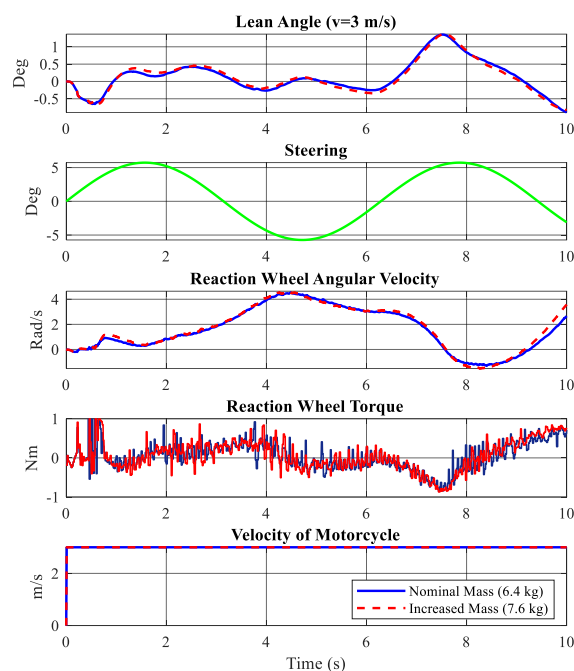
به منظور ارزیابی مقاومت سیستم در برابر اغتشاشات محیطی، یک دست‌انداز به عنوان نویز خارجی در مسیر حرکت دوچرخه اعمال شد و نتایج حاصل از دو الگوریتم DDPG و TD3 به ترتیب در شکل ۱۹ و ۲۰ ارائه شده‌اند. همان‌طور که مشاهده می‌شود، هر دو الگوریتم موفق شده‌اند پس از برخورد با دست‌انداز، زاویه غلتش را در محدوده‌ای پایدار (کمتر از ۴ درجه) نگه داشته و دوچرخه را با موفقیت به مقصد برسانند.

با این حال، تفاوت قابل توجهی در رفتار دو الگوریتم وجود دارد. در الگوریتم DDPG (شکل ۱۹)، سرعت زاویه‌ای چرخ عکس‌العمل پس از دفع اغتشاش در محدوده‌ای امن و کنترل‌شده باقی می‌ماند و روند واگرایی مشاهده نمی‌شود. در مقابل، در الگوریتم TD3 (شکل ۲۰)، سرعت زاویه‌ای چرخ عکس‌العمل دارای روند افزایشی پیوسته است که نشان‌دهنده خطر اشباع عملگر در صورت تداوم حرکت می‌باشد.



شکل ۲۲. ارزیابی مقاومت کنترل کننده در برابر نویز اندازه گیری و دفع موفقیت آمیز اغتشاشات حسگر

Figure 22. Evaluation of controller robustness against measurement noise and the successful rejection of sensor disturbances



شکل ۲۱. ارزیابی مقاومت سیستم در برابر عدم قطعیت جرمی (افزایش ۱۶ درصدی جرم) حین حرکت مارپیچ فرمان سینوسی

Figure 21. Evaluation of system robustness against mass uncertainty (16% mass increase) during a continuous sine-wave steering motion

۴. نتیجه گیری

در این پژوهش، رویکردی هوشمند برای کنترل تعادل دوچرخه‌ی خودم تعادل با بهره‌گیری از الگوریتم‌های یادگیری تقویتی DDPG و TD3 ارائه شد. در این ساختار، کنترل کننده‌ی PID تنها برای مدیریت زاویه‌ی فرمان و ناپایداری در حضور موانع به کار گرفته شد، در حالی که فرایند حفظ تعادل جانبی به صورت هوشمند و بدون نیاز به مدل دینامیکی دقیق توسط عامل یادگیرنده انجام گرفت که از نوآوری‌های اصلی پژوهش محسوب می‌شود. همچنین، ارزیابی‌های مقایسه‌ای با کنترل کننده‌ی کلاسیک بهینه نشان داد که کنترل کننده‌های مبتنی بر یادگیری تقویتی توانمندی بسیار بالاتری در غلبه بر اثرات غیرخطی سیستم داشته و به ویژه در حین عبور از مسیرهای منحنی، پایداری ایمن تری را با نوسانات کمتر تضمین می‌کنند.

نتایج کمی شبیه سازی نشان داد که الگوریتم DDPG پس از 3579 اپیزود به سیاست پایدار رسید و RMSE زاویه‌ی غلتش را در بازه‌ی 0 تا 18 ثانیه به 1.7 درجه رساند. در مقابل، الگوریتم TD3 با نیاز به 7537 اپیزود برای همگرایی، این مقدار را به 1.23

در آزمون دوم، به منظور بررسی پایداری سیستم در حضور خطاهای ذاتی و نویز حسگرهای اندازه گیری، یک نویز نوسانی با فرکانس بالا به سنسور زاویه غلتش اعمال گردید. نتایج شبیه سازی در شکل ۲۲ نشان می‌دهد که اگرچه نویز اندازه گیری منجر به بروز نوسانات فرکانس بالا در گشتاور چرخ عکس العمل شده است، اما ساختار یادگیری عمیق، کنترل کننده توانسته است این اغتشاشات را به طور مؤثر بی اثر کرده و از ناپایداری و سقوط دوچرخه جلوگیری کند. این نتایج گواه بر مقاومت بسیار بالای چارچوب پیشنهادی در برابر دینامیک‌های مدل نشده و نویزهای محیطی است.

درجه کاهش داد. DDPG با تعداد اپیزود کمتر و مصرف انرژی پایین تر، یادگیری سریع تر دارد و برای کاربردهای دارای محدودیت سخت افزاری مناسب تر است، در حالی که TD3 از نظر پایداری نهایی، کاهش خطای ماندگار و مقاومت در برابر اغتشاش عملکرد برتری ارائه می دهد.

لازم به ذکر است که این پژوهش در محیط شبیه سازی انجام شده و برخی محدودیت های پیاده سازی سخت افزاری نظیر اشباع عملگر چرخ عکس العمل (در صورت رسیدن سرعت زاویه ای به حد بیشینه) و مدل ساده شده ی موتور (دینامیک الکتریکی و اصطکاک مکانیکی) نیازمند بررسی های بیشتر در فاز عملی هستند. با این وجود، ارزیابی های تکمیلی مقاومت سیستم اثبات نمود که چارچوب کنترلی پیشنهادی از پایداری بسیار بالایی برخوردار است. به طور مشخص، نتایج ارائه شده در بخش ۳.۳ (شکل های ۲۱ و ۲۲) نشان داد که الگوریتم TD3 در هر دو سناریوی آزمون شده افزایش ۱۶ درصدی جرم کل سیستم و اعمال نویز فرکانس بالا به حسگر زاویه ی غلتش بدون هیچ گونه بازتنظیم پارامتری، تعادل دوچرخه را به طور کامل حفظ نمود و هیچ رخداد ناپایداری مشاهده نشد.

در تحقیقات آینده، می توان چارچوب کنترلی پیشنهادی را با تعمیم به سناریوهای دارای سرعت پیشروی متغیر و بررسی اثرات شتاب گیری و ترمزگیری طولی بر تعادل جانبی توسعه داد. همچنین بهره گیری از الگوریتم های یادگیری تقویتی پیشرفته برای تشخیص خودکار موانع و ناوبری هوشمند در محیط های پویا، استفاده از یادگیری تقویتی جهت تنظیم خودکار و برخط ضرایب کنترل کننده ی PID فرمان، و در نهایت پیاده سازی تجربی این چارچوب بر روی یک پلتفرم فیزیکی به منظور ارزیابی عملکرد الگوریتم ها در مواجهه با چالش های سخت افزاری دنیای واقعی، از مهم ترین مسیرهای توسعه ی این پژوهش به شمار می روند.

9. Zheng, Q., Tian, Y., Deng, Y., Zhu, X., Chen, Z. and Liang, B., 2023. Reinforcement learning-based control of single-track two-wheeled robots in narrow terrain. *Actuators*, 12(3), p.109. <https://doi.org/10.3390/act12030109>
10. Zhu, X., Deng, Y., Zheng, X., Zheng, Q., Liang, B. and Liu, Y., 2022. Online reinforcement-learning-based adaptive terminal sliding mode control for disturbed bicycle robots on a curved pavement. *Electronics*, 11(21), p.3495. <https://doi.org/10.3390/electronics11213495>
11. Jalali, A., Mohammadi, N. and Tayefi, M., 2024. Reinforcement learning and sliding mode hybrid controller for an enhanced inverted pendulum on cart system. *Modares Mechanical Engineering*, 26(4), pp.241–251 [In Persian]. <https://dorl.net/dor/20.1001.1.10275940.1404.26.4.24.1>
12. Vijaya Lakshmi, K. and Manimozhi, M., 2024. Implementation of DDPG-based reinforcement learning control for self-balancing motorcycle. *IEEE Access*, 12, pp.117271–117284. <https://doi.org/10.1109/ACCESS.2024.3447054>
13. Zhang, D., Qi, L. and Luan, W., 2024. Reinforcement-based collision avoidance strategy for autonomous vehicles to multiple two-wheelers at un-signalized obstructed intersections. In *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Kuching, Malaysia, pp.3193–3198. IEEE. <https://doi.org/10.1109/SMC54092.2024.10831019>
14. Lai, F., Hu, H. and Huang, C., 2024. Trajectory preview tracking control for self-balancing intelligent motorcycle utilizing front-wheel steering. *Applied System Innovation*, 7(6), p.108. <https://doi.org/10.3390/asi7060108>
15. Trung, N.K., Tham, N.T., Diep, T.T., Ngoc, 1. Kondás, P. and Kapitány, P., 2023. Balancing control of a motorcycle. *Pollack Periodica*, 18(2), pp.167–171. <https://doi.org/10.1556/606.2022.00612>
2. Badgujar, C. and Mohite, S., 2023. Design, analysis and implementation of control moment gyroscope (CMG) mechanism to self-balance a moped bike. *Materials Today: Proceedings*, 72, pp.1517–1523. <https://doi.org/10.1016/j.matpr.2022.09.380>
3. Tran, T., Ho, T., Nguyen, H., Le, N., Tran, V. and Tran, Q., 2024. PID control for balancing bike model using reaction wheel. *Journal of Fuzzy Systems and Control*, 2(2), pp.50–57. <https://doi.org/10.59247/jfsc.v2i2.188>
4. Chiu, C.H. and Wu, C.Y., 2020. Bicycle robot balance control based on a robust intelligent controller. *IEEE Access*, 8, pp.84837–84849. <https://doi.org/10.1109/ACCESS.2020.2992792>
5. Tsuchiya, M., Hara, S., Kimura, T. and Tsurumi, N., 2022. Robust control strategy for robotic motorcycle without falling down at low-speed driving. *International Journal of Automotive Engineering*, 13(4), pp.188–195. https://doi.org/10.20485/jsaeijae.13.4_188
6. Seekhao, P., Tungpimolrut, K. and Parnichkun, M., 2020. Development and control of a bicycle robot based on steering and pendulum balancing. *Mechatronics*, 69, p.102386. <https://doi.org/10.1016/j.mechatronics.2020.102386>
7. Horoub, M., Alzaydi, A. and Hanieh, A.A., 2023. Electric ducted fan controller and self-balancing system for two-wheeler motorbike. *Frontiers in Mechanical Engineering*, 9, p.1284879. <https://doi.org/10.3389/fmech.2023.1284879>
8. Cachaya, E., Perdomo, C., Suaza, D., Zarta, J. and Serrezuela, R., 2024. Control of autonomous motorcycles by means of trajectory tracking and balance stabilization. *Journal of Theoretical and Applied Information Technology*, 102(7), pp.3093–

22. Singh, R. and Bera, T.K., 2018. Obstacle avoidance of bicycle vehicle model using overwhelming controller. *Arabian Journal for Science and Engineering*, 43(9), pp.4821–4833. <https://doi.org/10.1007/s13369-018-3175-5>
23. Tayefi, M. and Geng, Z., 2018. Self-balancing controlled Lagrangian and geometric control of unmanned mobile robots. *Journal of Intelligent & Robotic Systems*, 90(1–2), pp.253–265. <https://doi.org/10.1007/s10846-017-0666-7>
24. Sadeghi, E., Karimi, J. and Sadati, S.H., 2017. UAV 3D robust guidance law design by fuzzy sliding mode approach. *Sharif Journal of Mechanical Engineering (SJME)*, 33.3(2), pp.13–25 [In Persian]. <https://doi.org/10.24200/j40.2018.10518.1297>
25. Mobarhani, A. and Taghirad, H.D., 2014. Mobile robot semantic mapping using histogram based frontier exploration. *Sharif Journal of Mechanical Engineering (SJME)*, 30.3(1), pp.115–124 [In Persian]. https://sjme.journals.sharif.edu/article_6267.html
26. Ashraf, N.M., Mostafa, R.R., Sakr, R.H. and Rashad, M.Z., 2021. Optimizing hyperparameters of deep reinforcement learning for autonomous driving based on whale optimization algorithm. *PLoS ONE*, 16(6), p.e0252754. <https://doi.org/10.1371/journal.pone.0252754>
- V.T.A. and Nguyen, H.Q., 2025. Research on the application of the model order reduction algorithm in designing a robust controller for the balance system of a self-balancing two-wheeled bicycle. *Engineering, Technology & Applied Science Research*, 15(1), pp.20484–20492. <https://doi.org/10.48084/etasr.9649>
16. De Vries, B., 2023. Modeling and implementation of a reaction wheel stabilization system for low speed balance of a cargo bicycle. *The Evolving Scholar*, Delft University of Technology, Netherlands. <https://doi.org/10.24404/63ff23b478f53b9c419075b9>
17. Hartono, R., Kim, S.B., Kalend, P.T.A., Baik, N.K. and Shin, K.J., 2023. Design of self-balancing model-size electrical motorbike robot using control moment gyroscope. *Sensors and Materials*, 35(1), pp.347–363. <https://doi.org/10.18494/SAM4168>
18. Vasilevski, D. and Pecioski, D., 2023. Design and implementation of self-balancing motorcycle. *Engineering Proceedings*, 41(1), p.109. <https://doi.org/10.55302/mesj23412670109v>
19. Tian, J., Ding, J., Tai, Y. and Ma, Z., 2020. Control of different-axis two-wheeled self-balancing vehicles. *IEEE Access*, 8, pp.158839–158851. <https://doi.org/10.1109/ACCESS.2020.3019538>
20. Wang, X. and Shenyang, X.Z., 2025. One-wheeled balance bike design based on LQR control algorithm. *Journal of Physics: Conference Series*, 3079, p.012059. <https://doi.org/10.1088/1742-6596/3079/1/012059>
21. Yi, J., Song, D., Levandowski, A. and Jayasuriya, S., 2006. Trajectory tracking and balance stabilization control of autonomous motorcycles. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Orlando, FL, USA, pp.2583–2589. IEEE. <https://doi.org/10.1109/ROBOT.2006.1642091>