Некоммерческое Акционерное Общество «Казахский Национальный Исследовательский Технический Университет имени К.И.Сатпаева»

Институт Автоматики и информационных технологий

Кафедра «Робототехники и технических средств автоматики»

Куанаев Арманжан Ахатұлы

Обратный маятник

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к дипломному проекту

Специальность 6B07111 – Робототехника и мехатроника

Алматы 2025 г

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН НЕКОММЕРЧЕСКОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ К.И. САТПАЕВА»

Институт Автоматики и информационных технологий

Кафедра «Робототехники и технических средств автоматики»

Куанаев Арманжан Ахатұлы

ДИПЛОМНЫЙ ПРОЕКТ

«Обратный маятник»

Специальность 6В07111 – Робототехника и мехатроника

Алматы 2025 г.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН НЕКОММЕРЧЕСКОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ К.И. САТПАЕВА»

SATBAYEV UNIVERSITY

Институт Автоматики и информационных технологий

Кафедра «Робототехники и технических средств автоматики»

ДОПУЩЕН К ЗАЩИТЕ
Заведующий кафедрой РТиТСА
кандидат технических наук
_____ Ожикенов К. А.
«03» ____06____ 2025 г.

ДИПЛОМНЫЙ ПРОЕКТ

Тема: «Обратный маятник»

по специальности 6В07111 – Робототехника и мехатроника

Выполнил Куанаев А. А.

Рецензент
PhD, Ассоциированный профессор
_____ Карымсакова Н. Т.
«30» май 2025 г.

Научный руководитель,
Старший преподаватель
_____ Рахметова П. М.
«30» май 2025 г.

Алматы 2025 г.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
НЕКОММЕРЧЕСКОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ К.И. САТПАЕВА»

SATBAYEV UNIVERSITY

Институт Автоматики и информационных технологий

Кафедра «Робототехники и технических средств автоматики»

6В07111 – Робототехника и мехатроника

ДОПУЩЕН К ЗАЩИТЕ
Заведующий кафедрой
РТиТСА кандидат технических наук
_____Ожикенов К. А.
«05» 06 2025 г.

ЗАДАНИЕ
на выполнение дипломной работы (проекта)

Студенту Куанаев Арманжан Ахатұлы
1. Тема: «Обратный маятник»
Утверждена приказом ректора_____ № _____от «__» _____ 202_ г.
2. Срок сдачи студентом законченной работы «__» _____202_ г.
3. Исходные данные к дипломному проекту (законы, литературные источники, лабораторно-производственные данные)
Теоретические материалы по Системам управления
Теоретические материалы по Обучению с подкреплением
Теоретические материалы по приложению Matlab
4. Перечень подлежащих разработке в дипломном проекте вопросов:
1) Изучить актуальность проблемы использования обучения с подкреплением в системах управления.
2) Исследовать систему управления маятника с использованием обучения с подкреплением.
3) Провести анализ существующих методов и алгоритмов в системах управления.
4) Настроить среду моделирования для проведения симуляции.
5) Осуществить визуализацию симуляции.
6) Провести анализ полученных данных из симуляции.
5. Перечень графического материала (чертежи, таблицы, диаграммы и т.д.)
Рисунки: 23
6. Перечень основной рекомендуемой литературы: *из* 4 *наименований* 4

ГРАФИК
подготовки дипломного проекта

Наименование разделов, перечень разрабатываемых вопросов	Сроки представления научному руководителю	Примечания
Исследовательская часть	05.06.2025	
Теоретическая часть	05.06.2025	
Практическая часть	08.06.2025	
Экспериментальная часть	08.06.2025	

Подписи
консультантов и нормоконтролера на законченный проект с указанием относящихся к ним разделов проекта

Наименование разделов	Консультанты, Ф.И.О. (уч. степень, звание)	Дата подписания	Подпись
Нормоконтролер	Ергебай Е. А.		
Основная часть			
Расчетная часть			
Программная часть			

Дата выдачи задания «05» июня 2025 г.

Научный руководитель: _____ Рахметова П. М.

Задание принял: _____ Куанаев А. А.

НЕКОММЕРЧЕСКОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ имени К.И. САТПАЕВА»

РЕЦЕНЗИЯ

Дипломной работы

студента по специальности 6B07111 – «Робототехника и мехатроника»

Куанаев Арманжан Ахатұлы

на тему: **«Обратный маятник»**

Дипломный проект Куанаев Арманжана выполнена по теме «Обратный маятник».

Цель работы – изучить и продемонстрировать эффективность алгоритмов обучения с подкреплением (RL) в стабилизации системы перевернутого маятника — широко признанной эталонной проблемы в теории управления. Целью данной работы является замена традиционных стратегий управления на основе моделей на подход, основанный на данных, способный обучаться оптимальным действиям посредством взаимодействия с окружающей средой.

В целом в работе с помощью приложения MATLAB Reinforcement Learning Designer агент обучения с подкреплением был эффективно обучен, визуализирован и тщательно протестирован. Агент достиг высокой производительности, успешно стабилизировав маятник в большинстве симуляций, тем самым подтвердив пригодность метода DQN для приложений управления в реальном времени в динамических и неопределенных средах.

Работа студента Куанаев Арманжан по теме «Обратный маятник» оцениваю на 75 , и считаю что студент Арманжан заслуживает присвоения академической степени бакалавра

Рецензент:
КазНУ имени аль-Фараби,
и.о. доцента кафедры искусственного интеллекта
и Big Data, PhD Карымкасова Н.Т
«___»_____ 2025 г.

АҢДАТПА

Бұл жұмыста «Басқару жүйелеріндегі күшейтетін оқыту» деп аталатын күшейтетін оқыту (RL) зерттелді. Автоматты басқару теориясының (ТАС) классикалық есебі «Кері маятник» динамизміне, хаосына және болжауға болмайтындығына байланысты таңдалды және моделі бақылаушы ретінде әрекет ететін Deep Q-Network (DQN) алгоритмін Reinforcement Learning (RL) пайдаланып оқытылды, оның міндеті кері маятникті тұрақтандыру және оны минималды оң позицияда ұстау. Тәжірибе көрсеткендей, инновациялық технологияларды қосымша оқыту әдісі ретінде пайдалану жоғары динамикалық, жоғары хаотикалық жүйелерді басқарумен сәтті күреседі. Мұндай әдісті болжау мүмкін еместігіне байланысты жүйенің математикалық моделін құру көп еңбекті қажет ететін жүйелерде, мысалы, гуманоид робот, автокөлікті жүргізу, бағдаршамдарды басқару арқылы автомобильдер ағынын оңтайландыру сияқты жүйелерде және т.б.

АННОТАЦИЯ

В этой работе было сделано исследование Reinforcement learning (RL) так называемый Обучение с подкреплением в Системах управления (Control Systems). Была выбрана классическая проблема из Теории Автоматического Управления (ТАУ) «Обратный маятник» из-за его динамичности, хаотичности и непредсказуемости и была обучена модель с использованием алгоритма Deep Q-Network (DQN) Обучения с подкрепления (RL) выступающий виде контроллера чья задача стабилизировать обратный маятник и держать его в вертикальном положении с минимальными действиями. Опыт показал, что использование инновационных технологии как метод Обучения с подкреплением (RL) успешно справляется с управлением высоко динамичными, высоко хаотичными системами. Такой метод может быть использован в системах, где трудоемко построить математическую модель системы из-за его непредсказуемости, например в системах как робот гуманоид, вождение автомобиля, оптимизация потока машин путем контроля светофоров и т.д.

ABSTRACT

In this paper, a study of Reinforcement learning (RL) was made, the so-called Reinforcement Learning in Control Systems. A classic problem from the Theory of Automatic Control (TAC) "Inverse Pendulum" was chosen due to its dynamism, chaos and unpredictability and a model was trained using the Deep Q-Network (DQN) algorithm of Reinforcement Learning (RL) acting as a controller whose task is to stabilize the inverse pendulum and keep it in an upright position with minimal actions. Experience has shown that the use of innovative technologies as a method of Reinforcement Learning (RL) successfully copes with the control of highly dynamic, highly chaotic systems. Such a method can be used in systems where it is labor-intensive to build a mathematical model of the system due to its unpredictability, for example in systems such as a humanoid robot, car driving, optimization of the flow of cars by controlling traffic lights, etc.

Перечень принятых сокращений

RL – Reinforcement learning
ТАУ – Теория автоматического управления
DQN – Deep Q-Network
PID – Proportional-integral-derivative controller
LQR – Linear Quadratic Regulator
ПИД - Пропорционально-интегрально-дифференцирующий регулятор
PPO - Proximal policy optimization
A2C – Advantage Actor-Critic
A3C - Asynchronous Advantage Actor Critic
DDPG - Deep Deterministic Policy Gradient
SAC – Soft Actor-Critic

СОДЕРЖАНИЕ

Введение 8
1.1 Структура системы управления 12
1.2 Обратный маятник 14
1.3 Алгоритмы Reinforcement Learning (RL) 15
1.4 Разделение Machine Learning: unsupervised, supervised, reinforcement learning 17
2 Математическое описание окружающей среды (тележки с маятником) 21
2.1 Структура обучения с подкреплением 23
2.2 Алгоритм Deep Q-Network (DQN) 24
2.3 Математическое описание Q-Network (QN) 24
2.4 Математическое описание Deep Q-Network (DQN) 25
2.5 Процедура обучения 26
2.6 Реализация симуляции и метрики симуляции 27
3 Результаты 29
3.1 Обучение модели RL 29
3.2 Визуализация 33
3.3 Анализ производительности модели 35
3.4 Результаты 41
3.5 Будущие улучшения 42
Заключение 43
Ссылки на литературу 45
Приложение А 46

Введение

Системы управления играют важную роль в современном технологическом ландшафте, обеспечивая стабильность и надежную работу в изначально нестабильных системах. Перевернутый маятник, эталонная задача в области инженерии управления, символизирует сложности, связанные с нелинейной динамикой, нестабильностью и чувствительностью к возмущениям. Его актуальность распространяется на практические приложения, такие как управление ориентацией космических аппаратов, роботизированные системы балансировки и персональные транспортные средства.

Традиционные стратегии управления, включая ПИД и классические методы пространства состояний, часто сталкиваются с ограничениями в условиях неопределенности и нелинейного поведения. Недавно обучение с подкреплением (RL) — парадигма машинного обучения, вдохновленная поведенческой психологией — стала многообещающей альтернативой из-за присущей ей адаптивности и надежности в динамических средах.

Эта работа решает проблему стабилизации перевернутого маятника с использованием методов обучения с подкреплением. Основные цели — применять, анализировать и сравнивать различные методы RL, такие как алгоритмы Deep Q-Network (DQN) и Proximal Policy Optimization (PPO). Эффективность и стабильность этих методов будут оцениваться посредством строгого вычислительного моделирования, фокусирующегося на реакции системы, отбрасывании помех и показателях производительности.

Методологический подход включает в себя вычислительное моделирование с использованием среды MATLAB. Хорошо зарекомендовавшие себя алгоритмы RL будут систематически внедряться, обучаться и тестироваться на этой платформе. Результаты будут сравниваться на основе ключевых показателей производительности, включая точность стабилизации, устойчивость к помехам и алгоритмическую эффективность. Интеграция MATLAB облегчает подробный анализ динамики системы, предоставляя строгую среду проверки.

Актуальность

В нынешнюю эпоху быстрого технологического прогресса интеграция интеллектуальных систем управления становится все более важной в различных областях техники, включая робототехнику, аэрокосмическую промышленность и автономные системы. Перевернутый маятник, классическая задача в теории управления, служит идеальным эталоном для тестирования и проверки передовых стратегий управления из-за присущей ему нестабильности и нелинейной динамики. Обучение с подкреплением (RL), подмножество машинного обучения, привлекло значительное внимание своей способностью изучать оптимальные политики управления посредством проб и ошибок во взаимодействии с окружающей средой, не требуя явных системных моделей. Актуальность этого исследования заключается в демонстрации потенциала алгоритмов RL, таких как Deep Q-Networks (DQN), для решения сложных задач управления в реальном времени и неопределенных условиях. Более того, исследование RL для стабилизации маятника обеспечивает фундаментальную основу для расширения таких подходов на реальные приложения, включая гуманоидных роботов, самобалансирующиеся транспортные средства и системы точного наведения, тем самым соответствуя растущему спросу на автономные и адаптивные решения управления.

Новизна

Новизна этой работы заключается в применении современных алгоритмов обучения с подкреплением (RL), в частности, Deep Q-Networks (DQN), к классической задаче управления стабилизацией перевернутого маятника. Хотя традиционные методы управления, такие как PID, LQR или методы пространства состояний, уже давно используются для этой задачи, они, как правило, требуют точного математического моделирования и ручной настройки, что ограничивает их адаптивность к изменяющейся динамике и возмущениям. Напротив, подход на основе RL, используемый в этом исследовании, позволяет агенту автономно изучать оптимальную политику управления посредством взаимодействия с окружающей средой, не полагаясь на предварительные знания уравнений системы. Один из примеров на рисунке 1.1. Эта методология, основанная на

данных, представляет собой гибкую масштабируемую структуру, способную обрабатывать нелинейности и неопределенности более эффективно, чем традиционные методы. Более того, интеграция нейронных сетей в DQN представляет собой значительный шаг к соединению инженерии управления и искусственного интеллекта, знаменуя собой переход к системам управления на основе обучения с потенциалом для внедрения в реальном времени и обобщения на другие нестабильные или недостаточно активируемые системы.

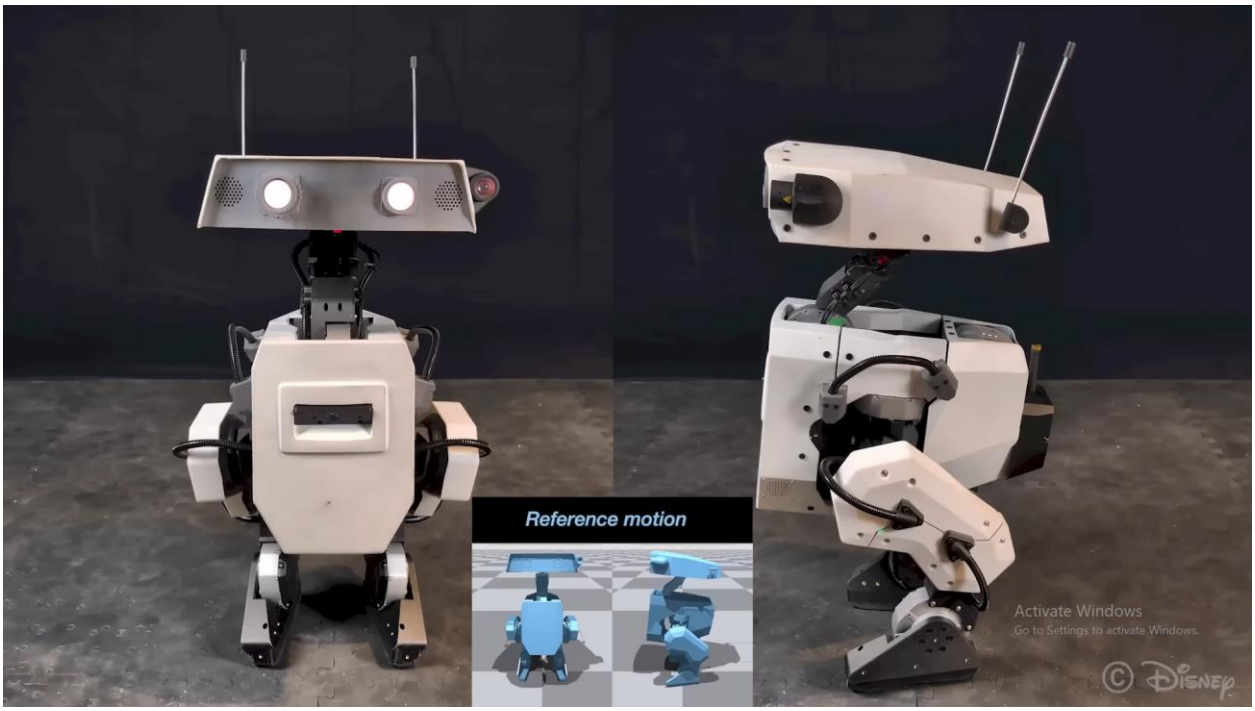


Рисунок 1.1 – Робот бипедал на основе обучения с подкреплением [8]

Цели и задачи

Основная цель данного исследования — изучить и продемонстрировать эффективность алгоритмов обучения с подкреплением (RL) в стабилизации системы перевернутого маятника — широко признанной эталонной проблемы в теории управления. Целью данной работы является замена традиционных стратегий управления на основе моделей на подход, основанный на данных, способный обучаться оптимальным действиям посредством взаимодействия с

окружающей средой. Для достижения этой цели в исследовании поставлены следующие цели:

1. Реализовать и обучить агента Deep Q-Network (DQN) для балансировки маятника в вертикальном положении.
2. Проанализировать эффективность обучения с точки зрения устойчивости, скорости сходимости и надежности.
3. Изучить потенциал для реализации в реальном времени и обобщения обученной модели для аналогичных задач нелинейного управления.

Метод исследования

В этом исследовании используется экспериментальный метод исследования на основе моделирования, объединяющий теорию управления, машинное обучение и системное моделирование. Система перевернутого маятника моделируется с использованием MATLAB для точного представления ее нелинейной динамики и нестабильности. Глубокая Q-сеть (DQN), алгоритм обучения с подкреплением на основе значений, реализована для обучения агента, который учится балансировать маятник, взаимодействуя с моделируемой средой. Процесс обучения включает определение пространств состояний и действий, разработку подходящей функции вознаграждения и настройку архитектуры нейронной сети, используемой для аппроксимации Q-значений. Метрики производительности, такие как кумулятивное вознаграждение, количество эпизодов до сходимости и стабильность с течением времени, отслеживаются на протяжении всего процесса обучения. Чтобы оценить эффективность контроллера RL, поведение обученного агента сравнивается с традиционными подходами к управлению, такими как PID или LQR, в той же структуре моделирования. Этот методологический подход обеспечивает контролируемую настройку для изучения возможностей RL в решении задач управления в реальном времени без опоры на явные модели системы.

1.1 Структура системы управления

Системы управления являются основополагающими компонентами в инженерии, используемыми для управления, регулирования и направления поведения динамических систем. Их основная функция заключается в обеспечении того, чтобы выход системы следовал желаемому опорному значению или заданному значению, несмотря на наличие внутренних или внешних возмущений. Типичная система управления состоит из установки (системы, которой нужно управлять), датчиков для измерения выходных данных, контроллеров для вычисления корректирующих действий и исполнительных механизмов для применения этих действий. Системы управления в целом подразделяются на системы с открытым и закрытым контуром (обратная связь). Системы с закрытым контуром чаще встречаются в современных приложениях, поскольку они непрерывно контролируют выходной сигнал и корректируют входной сигнал для минимизации ошибок. Системы управления играют важную роль в различных областях, включая робототехнику, аэрокосмическую, автомобильную, обрабатывающую и перерабатывающую промышленность, обеспечивая автоматизацию, точность и стабильность в сложных задачах. На рисунке 2.2, можно наблюдать классическую схему системы управления.

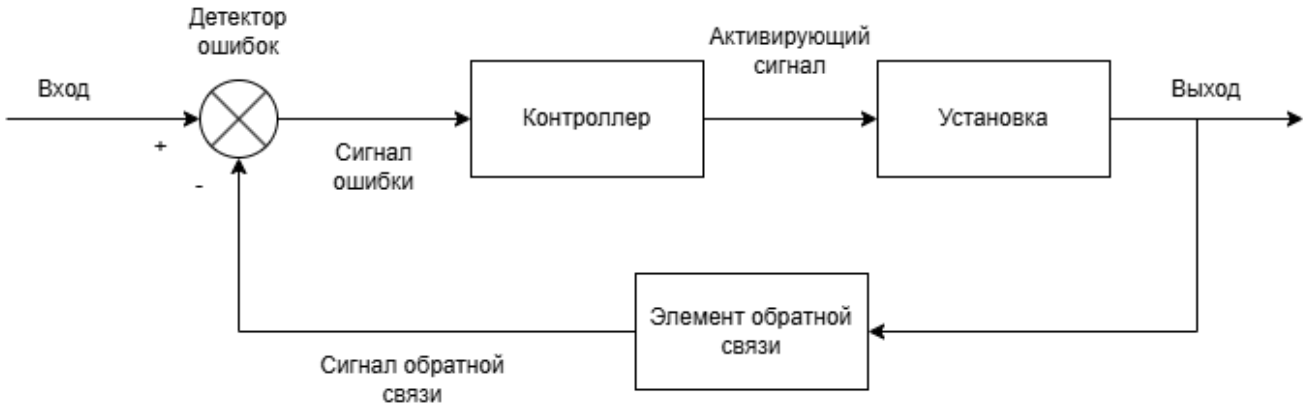


Рисунок 2.1 – Классическая система управления [4]

Задача перевернутого маятника долгое время служила эталоном в теории управления и робототехнике из-за присущей ей нестабильности и проблем, которые она представляет при управлении в реальном времени. Традиционные

методы управления, такие как пропорционально-интегрально-дифференциальные (ПИД) регуляторы и линейные квадратичные регуляторы (LQR), широко использовались для стабилизации маятника путем выведения оптимальных стратегий управления на основе математических моделей динамики системы. Однако эти классические подходы часто опираются на линеаризованные модели, которые могут не охватывать нелинейности системы при больших возмущениях или неопределенностях.

В последние годы обучение с подкреплением (RL) стало мощной альтернативой для задач управления, позволяя агентам изучать оптимальные политики напрямую через взаимодействие с окружающей средой. В отличие от классических методов управления, RL не требует явного моделирования системы и может обрабатывать многомерные пространства состояний и сложную динамику. Ключевые алгоритмы, такие как Q-learning, методы градиента политики и фреймворки Actor-Critic, продемонстрировали успех в различных приложениях управления и робототехники.

Заметным достижением в RL является введение глубоких Q-сетей (DQN) Мнихом (2015) [13], которые объединяют Q-обучение с глубокими нейронными сетями для аппроксимации функции действия-значения в многомерных пространствах. Этот подход значительно улучшил применимость RL к сложным задачам управления, включая задачи с непрерывными состояниями и дискретными действиями. DQN стабилизирует обучение с использованием воспроизведения опыта и целевых сетей, решая проблемы корреляции данных и нестационарных целевых значений.

В нескольких исследованиях изучалось применение RL к стабилизации перевернутого маятника. Например, Бен и др. (2024) [5] применили Q-обучение для стабилизации моделируемой системы тележка-шест, продемонстрировав осуществимость подходов, основанных на обучении. Совсем недавно исследователи расширили эту работу, используя глубокие алгоритмы RL, такие как DQN и Double DQN, демонстрируя улучшенную стабильность обучения и сходимость как в симуляциях, так и в реальных сценариях. Эти исследования подчеркивают способность DQN изучать политики управления без предварительного знания динамики системы и адаптироваться к изменяющимся условиям.

Более того, сравнение контроллеров на основе RL и традиционных методов показало, что RL может достигать конкурентоспособной или превосходной

производительности, особенно в средах с неопределенностями модели или шумом. Однако методы управления на основе RL часто требуют значительного времени обучения и вычислительных ресурсов, которые остаются активными областями исследований для повышения эффективности и надежности выборки.

Подводя итог, литература показывает, что обучение с подкреплением, и, в частности, глубокие Q-сети, предлагают многообещающий подход к решению проблемы стабилизации перевернутого маятника. Используя глубокие нейронные сети для аппроксимации оптимальных политик, DQN может обрабатывать нелинейную динамику и изучать эффективные стратегии управления посредством взаимодействия с окружающей средой. Этот проект строится на этих основах, применяя DQN для стабилизации системы перевернутого маятника и исследуя ее производительность по сравнению с традиционными подходами к управлению.

1.2 Обратный маятник

Обратный маятник является классической и широко изучаемой проблемой в области техники управления из-за присущей ему нестабильности и нелинейной динамики. Он служит идеальным эталоном для тестирования и оценки стратегий управления, поскольку даже небольшие отклонения от вертикального положения приводят к быстрому расхождению без корректирующего воздействия. Несмотря на простоту конструкции, управление перевернутым маятником требует точных и своевременных действий, что делает его ценной моделью для разработки и проверки передовых методов управления. Более того, он имеет схожие характеристики со многими реальными системами, такими как двуногие роботы, самобалансирующиеся транспортные средства, системы наведения ракет и стабилизация ракет. Изучение перевернутого маятника обеспечивает глубокое понимание управления с обратной связью, динамики системы и производительности алгоритма в условиях неопределенности, что делает его важным образовательным и исследовательским инструментом как в классической, так и в современной теории управления. На рисунке 2.2 можно увидеть эталонную классическую проблему из теории автоматического управления - обратный маятник.

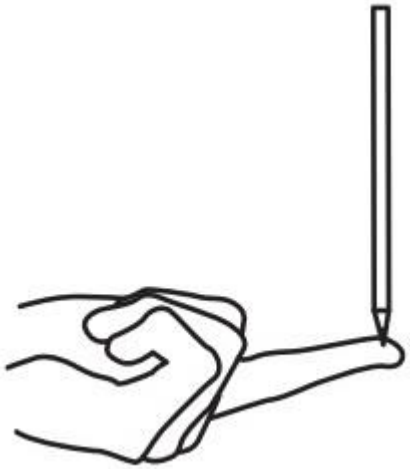


Рисунок 2.2 – Эталонная проблема теории управления обратный маятник [3]

1.3 Алгоритмы Reinforcement Learning (RL)

Обучение с подкреплением (RL) — это тип машинного обучения, в котором агент учится принимать решения, взаимодействуя со средой для достижения определенной цели. В отличие от контролируемого обучения, которое опирается на маркированные данные, RL основано на процессе проб и ошибок. Агент наблюдает за текущим состоянием среды, выполняет действие и получает обратную связь в виде вознаграждения или штрафа. Со временем агент учится выбирать действия, которые максимизируют кумулятивное вознаграждение. Эта структура обучения хорошо подходит для задач управления, где среда является динамичной и сложной, и где оптимальные действия заранее неизвестны. Обучение с подкреплением привлекло значительное внимание благодаря своему успеху в робототехнике, играх, автономных системах и задачах управления в реальном времени, что делает его мощным инструментом для разработки интеллектуальных и адаптивных контроллеров. На рисунке 2.3 проиллюстрировано базовая схема работы модели на базе обучения с подкреплением.

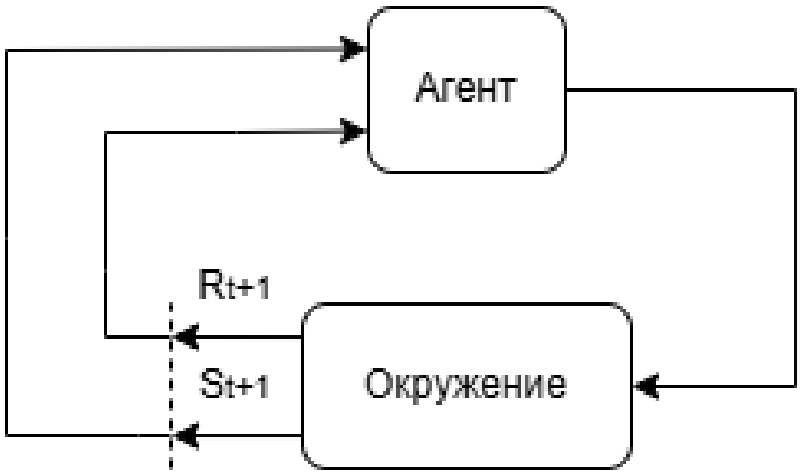


Рисунок 2.3 – Базовая схема работы модели на базе обучения с подкреплением [12]

Алгоритмы обучения с подкреплением (RL) можно в целом разделить на три основных типа: методы на основе ценностей, политики и актор-критик. Алгоритмы на основе ценностей, такие как Q-Learning и Deep Q-Networks (DQN), фокусируются на оценке ожидаемого совокупного вознаграждения (ценности) для каждого действия в заданном состоянии и выбирают действие с наивысшей ценностью. Рисунок 2.4 показывает категоризацию различных методов обучения с подкреплением. Алгоритмы на основе политик, такие как REINFORCE и Proximal Policy Optimization (PPO), напрямую изучают политику — отображение состояний в действия — не полагаясь на оценку ценности. Эти методы хорошо подходят для задач с непрерывными пространствами действий. Алгоритмы актор-критик объединяют оба подхода, используя актор для выбора действий и критика для их оценки, что приводит к более стабильному и эффективному обучению. Примерами являются Advantage Actor-Critic (A2C) и Deep Deterministic Policy Gradient (DDPG). Выбор алгоритма зависит от сложности среды, характера пространства действия (дискретное или непрерывное) и желаемого компромисса между исследованием и эксплуатацией. Эти алгоритмы формируют основу современных приложений RL в робототехнике, системах управления, играх и автономном принятии решений.

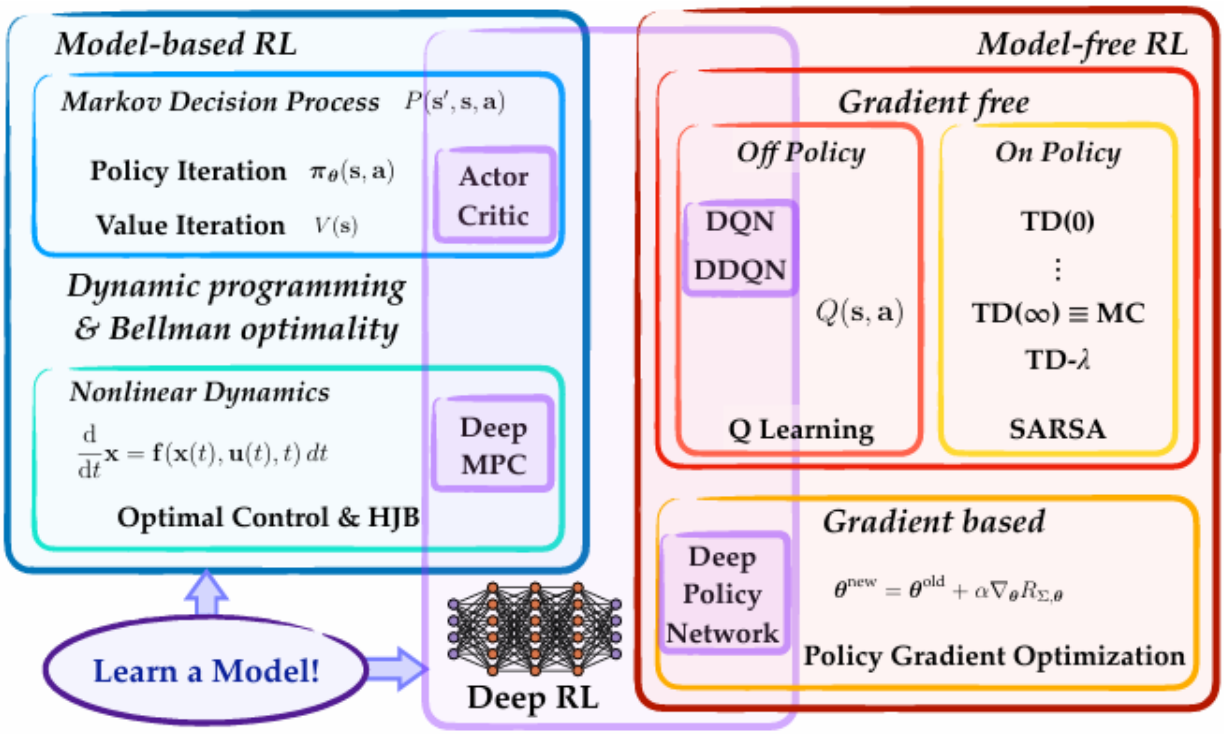


Рисунок 2.4 – Категоризация методов обучения с подкреплением [1]

1.4 Разделение Machine Learning: unsupervised, supervised, reinforcement learning

Машинное обучение обычно делится на три основные категории: контролируемое обучение, неконтролируемое обучение и обучение с подкреплением. При контролируемом обучении алгоритм обучается на маркированных данных, где входные данные сопоставляются с правильными выходными данными. Цель состоит в том, чтобы изучить функцию сопоставления, которая может предсказывать результаты для новых, невиданных данных. Этот подход обычно используется в таких задачах, как классификация изображений, обнаружение спама и проблемы регрессии. Напротив, неконтролируемое обучение имеет дело с немаркированными данными. Алгоритм пытается найти скрытые закономерности или структуры в данных, такие как группировка схожих элементов (кластеризация) или сокращение измерений данных. Его часто используют для анализа данных, обнаружения аномалий и сегментации клиентов. Обучение с подкреплением — это другая

парадигма, в которой агент обучается, взаимодействуя со средой. Он получает вознаграждения или штрафы на основе своих действий и стремится максимизировать общее вознаграждение с течением времени. Этот метод особенно полезен для задач управления, робототехники и игр, где решения должны приниматься последовательно, а результаты зависят как от текущих, так и от прошлых действий.

Традиционные системы управления обычно проектируются с использованием точных математических моделей физической системы. Эти модели часто опираются на предположения и упрощения, такие как линеаризация вокруг рабочей точки, которая может пренебречь важными нелинейностями и динамическими взаимодействиями. В результате классические контроллеры, такие как ПИД или LQR, могут хорошо работать в идеальных условиях, но испытывать трудности при столкновении с неопределенностью, не моделированной динамикой или изменяющейся средой. Обучение с подкреплением (RL) предлагает принципиально иной подход: оно не требует явной модели системы. Вместо этого агент RL изучает оптимальные стратегии управления посредством взаимодействия с окружающей средой, адаптируя свое поведение на основе опыта и обратной связи. На рисунке 2.5 показан принцип работы обучения с подкреплением. Это делает RL особенно привлекательным для сложных, нелинейных или плохо изученных систем, где традиционное моделирование затруднено или неточно. Постоянно улучшая свою политику методом проб и ошибок, RL может достичь надежного и гибкого управления, что делает его мощным инструментом для следующего поколения интеллектуальных и автономных систем.

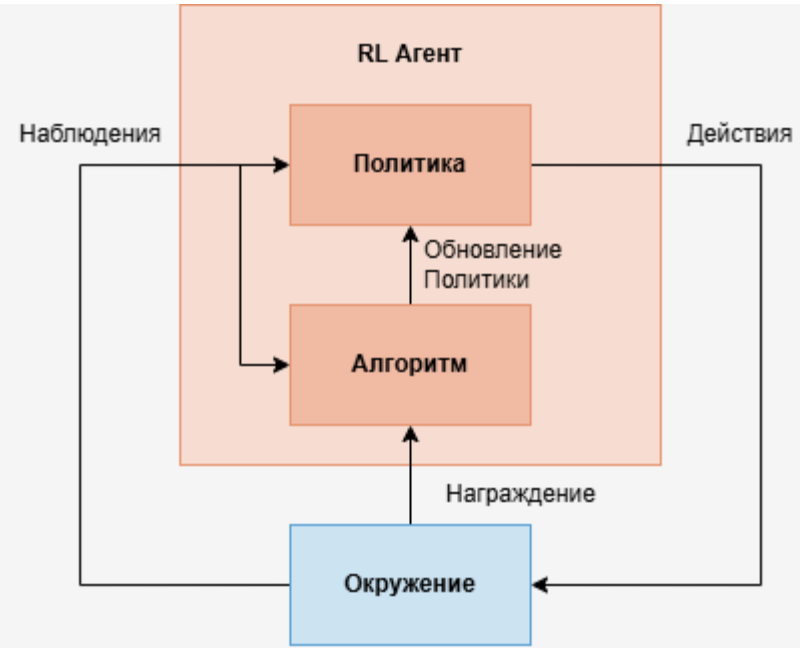


Рисунок 2.5 – Принцип работы обучения с подкреплением

Алгоритм Deep Q-Network

Deep Q-Network (DQN) — это алгоритм обучения с подкреплением на основе значений, который объединяет Q-Learning с глубокими нейронными сетями для обработки сред с большими или непрерывными пространствами состояний. В традиционном Q-Learning Q-таблица используется для хранения ожидаемых вознаграждений для каждой пары состояние-действие, что становится непрактичным, когда количество состояний очень велико. DQN устраняет это ограничение, используя нейронную сеть, называемую Q-сетью, для аппроксимации Q-значений. Входом в сеть является текущее состояние, а выходом — оценочное значение для каждого возможного действия. Во время обучения сеть обновляется с использованием воспроизведения опыта и целевой сети для улучшения стабильности и сходимости. DQN позволяет агентам изучать эффективные политики управления непосредственно из многомерных сенсорных входов, таких как изображения или сложные симуляции. Он успешно применяется в различных областях, включая видеоигры, управление роботами и автономные системы, что делает его основополагающим алгоритмом в глубоком

обучении с подкреплением. На рисунке 2.6 показана принцип работы обучения с подкреплением на базе алгоритма DQN.

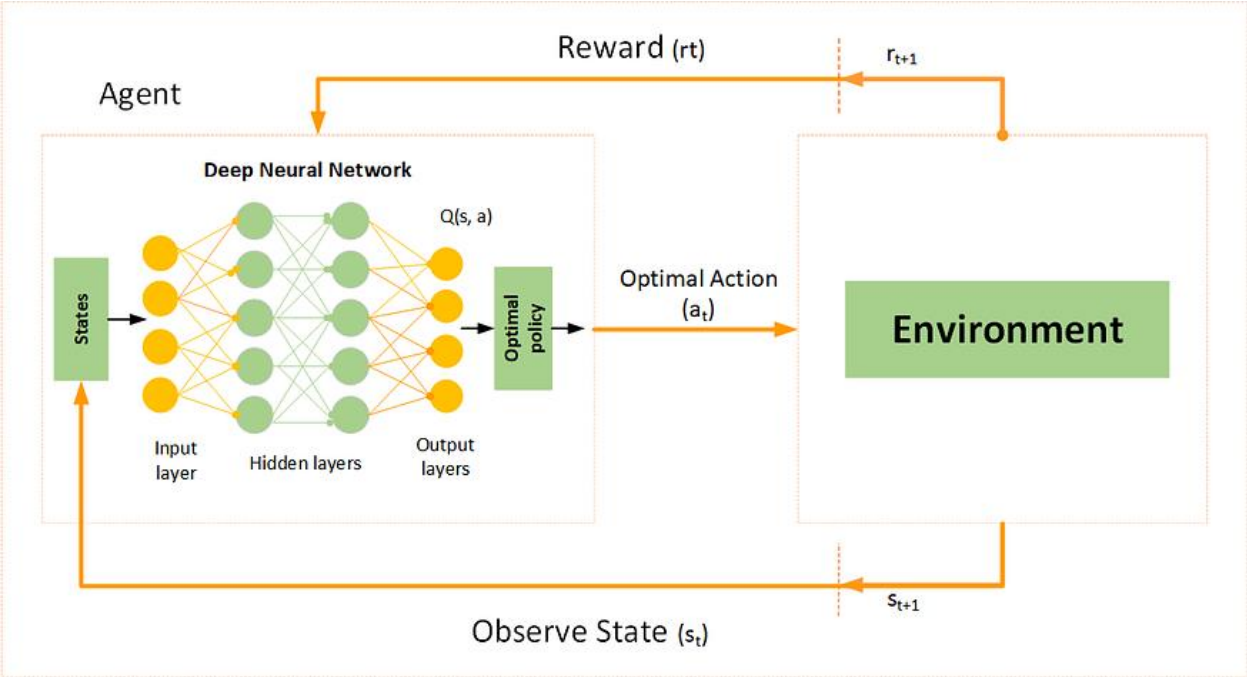


Рисунок 2.6 - Принцип работы обучения с подкреплением на базе алгоритма DQN [2]

2. Математическое описание Обратного маятника

В этой главе описывается методологический подход, принятый для достижения стабилизации системы перевернутого маятника с использованием Deep Q-Network (DQN), алгоритма обучения с подкреплением. Методология охватывает следующие основные этапы: определение среды обучения с подкреплением, выбор и настройка алгоритма DQN, обучение нейронной сети и реализация среды моделирования с использованием MATLAB.

Задача перевернутого маятника моделируется как классическая задача управления, в которой маятник должен быть сбалансирован в вертикальном положении путем управления положением и скоростью поддерживающей тележки. Среда определяется следующими переменными состояния:

– Угол маятника (θ)

– Угловая скорость (θ’)

– Положение тележки (x)

– Скорость тележки (x’)

Пространство действий дискретно, включает команды для перемещения тележки влево или вправо, обеспечивая достаточный крутящий момент для поддержания равновесия.

Перевернутый маятник, установленный на тележке, является классическим примером нелинейной динамической системы, широко используемой для демонстрации и проверки различных методологий управления, включая обучение с подкреплением. Система состоит из тележки, которая движется горизонтально, и жесткого маятника, шарнирно прикрепленного к тележке, который может свободно вращаться вокруг этой оси.

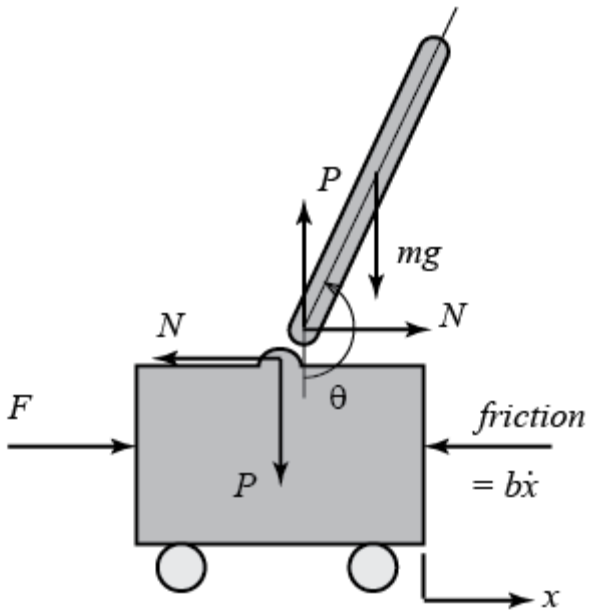


Рисунок 3.1 – Схема математической модели тележки с маятником [7]

Описание системы динамического поведения системы обратного маятника-тележки регулируется законами Ньютона и может быть охарактеризовано двумя связанными дифференциальными уравнениями, представляющими поступательное движение тележки и вращательное движение маятника. На рисунке 3.1 показана схема для обратного маятника.

Уравнения движения для вывода уравнений движения рассмотрим следующие параметры:

– Масса тележки (kg)

– Масса маятника (kg)

– Длина до центра масс маятника (m)

– Угловое смещение маятника от вертикали (rad)

– Горизонтальное смещение тележки (m)

– Ускорение под действием силы тяжести (g)

– Внешняя горизонтальная сила, приложенная к тележке (N)

Используя механику Лагранжа или механику Ньютона, получаем следующие нелинейные дифференциальные уравнения:

$$(M+m)\ddot{x} + ml\left(\ddot{\theta}\cos\theta - \dot{\theta}^2\sin\theta\right) = F \quad (1)$$

$$ml\ddot{x}\cos\theta + ml^2\ddot{\theta} - mgl\sin\theta = 0 \quad (2)$$

Эти уравнения можно преобразовать в более удобную матричную форму для приложений управления следующим образом:

$$\begin{bmatrix} M+m & ml\cos\theta \\ ml\cos\theta & ml^2 \end{bmatrix} + \begin{bmatrix} \ddot{x} \\ \ddot{\theta} \end{bmatrix} = \begin{bmatrix} F + ml\dot{\theta}^2 \sin\theta \\ mgl\sin\theta \end{bmatrix} \quad (3)$$

Линеаризованная модель для практического проектирования систем управления нелинейные уравнения часто линеаризуются вокруг точки равновесия x=0, θ=0, x˙=0 и θ˙=0. Полученная линеаризованная модель, действительная для малых угловых смещений, определяется следующим образом:

$$(M+m)\ddot{x} + ml\ddot{\theta} = F \quad (4)$$

$$ml\ddot{x} + ml^2\ddot{\theta} - mgl\theta = 0 \quad (5)$$

Эта линеаризованная система позволяет упростить и повысить эффективность проектирования и анализа различных стратегий управления, включая классические ПИД-регуляторы, контроллеры в пространстве состояний и алгоритмы обучения с подкреплением.

2.1 Структура обучения с подкреплением

Структура обучения с подкреплением для этого проекта состоит из трех основных компонентов:

– Агент: контроллер на основе нейронной сети, использующий алгоритм Deep Q-Network.

– Среда: смоделированный сценарий перевернутого маятника, созданный с помощью MATLAB.

– Структура вознаграждения: агент получает положительные вознаграждения за поддержание маятника в вертикальном положении (угол, близкий к нулю) и отрицательные вознаграждения, пропорциональные отклонению от вертикального равновесия, чрезмерному смещению тележки или падению маятника за пределы определенного порогового угла.

2.2 Алгоритм Deep Q-Network (DQN)

DQN был выбран из-за его эффективности в обработке дискретных пространств действий и возможности аппроксимации сложных политик с помощью нейронных сетей. Основные элементы алгоритма DQN, реализованные в этом исследовании, включают:

– Воспроизведение опыта: используется для хранения прошлого опыта и случайной выборки мини-пакетов для стабилизации обучения нейронной сети.

– Политика эпсилон-жадности: балансирует исследование и эксплуатацию во время обучения, регулируя скорость исследования (ε), постепенно уменьшаясь от начального высокого значения до минимального порога.

– Целевая сеть: периодически обновляемые веса из первичной Q-сети для стабилизации обучения и сходимости.

Архитектура сети состоит из полностью связанных слоев, оптимизированных с помощью итеративного экспериментирования с различной глубиной, количеством нейронов и параметрами обучения, что обеспечивает оптимальную эффективность и производительность обучения.

2.3Математическое описание Deep Q-Network (DQN)

Q-Learning построен на концепции Q-функции, также известной как функция ценности состояния-действия. Для заданной политики π, Q-функция $Q^{\pi}(s, a)$ оценивает ожидаемую кумулятивную награду при выполнении действия α в состоянии s, а затем при следовании политике π после этого. Цель состоит в том, чтобы определить оптимальную Q-функцию $Q^{*}(s, a)$, которая представляет собой максимальную ожидаемую отдачу, достижимую из состояния s путем выполнения действия α и затем оптимального действия на последующих этапах.

Следующая оптимальная Q-функция удовлетворяет уравнению оптимальности Беллмана:

$$Q^{*}(s, a) = \sum \left[r + \gamma \max_{a'} Q^{*}(s', a') \right] \tag{6}$$

В этом уравнении оптимальная отдача от пары состояние-действие (s, a) равна ожидаемому значению немедленного вознаграждения r плюс

дисконтированное максимальное значение Q следующего состояния s', предполагая, что оптимальные действия будут предприняты впоследствии. Фактор дисконтирования γ (с $0 \leq \gamma < 1$) гарантирует, что будущие вознаграждения будут соответствующим образом взвешены. Ожидание учитывает все возможные переходы и связанные с ними вознаграждения.

Итеративный процесс Q-Learning использует это уравнение Беллмана для обновления Q-функции:

$$Q_{i+1}(s,a) \leftarrow \sum \left[r + \gamma \max_{a'} Q_i(s',a') \right] \tag{7}$$

При повторных обновлениях Q-функция Q_i сходится к оптимальной Q-функции Q^*, когда число итераций i стремится к бесконечности. Этот принцип лежит в основе более продвинутых алгоритмов, таких как Deep Q-Network (DQN), где нейронная сеть используется для аппроксимации Q-функции.

Во многих реальных сценариях хранение Q-значений для каждой возможной пары состояние-действие в таблице нецелесообразно из-за размера пространства состояние-действие. Вместо этого Deep Q-Learning аппроксимирует Q-функцию с помощью параметризованной модели, обычно нейронной сети с параметрами θ, такой что $Q(s,a;\,\theta) \approx Q^*(s,a)$. Это приближение обучается путем минимизации функции потерь на каждом шаге i:

$$L_i(\theta_i) = \sum\nolimits_{s,a,r,s' \sim \rho(.)} [(y_i - Q(s,a;\,\theta_i))^2] \tag{8}$$

Где

$$y_i = r + \gamma \max_{a'} Q(s',a';\,\theta_{i-1}) \tag{9}$$

Здесь y_i называется целевой разницей во времени (TD), а разность $y_i - Q$ называется ошибкой TD. Ожидание берется по распределению ρ переходов $\{s,a,r,s'\}$, собранных посредством взаимодействия с окружающей средой.

Важно отметить, что целевые Q-значения вычисляются с использованием более ранней версии сети с параметрами θ_{i-1}, которые остаются постоянными во время обновления. В практических реализациях для этой цели используется

отдельная сеть, известная как целевая сеть. Она обновляется реже, чем основная сеть, чтобы улучшить стабильность во время обучения.

DQN — это алгоритм обучения вне политики. Он изучает значение оптимальной политики — выбирая действие $a = \max_a Q(s, a;\ \theta)$ — и при этом собирает опыт, используя другую политику поведения. Обычно эта политика поведения — ε-greedy, что означает, что агент выбирает наиболее известное действие большую часть времени (с вероятностью $1 - \varepsilon$) и исследует случайным образом (с вероятностью ε), чтобы в достаточной степени исследовать окружающую среду.

2.4 Процедура обучения

Обучение выполнялось итеративно с помощью встроенного в MATLAB инструментария обучения с подкреплением:

1. Инициализация параметров Q-сети и целевой сети.
2. Выполнение нескольких эпизодов обучения, в течение каждого из которых агент взаимодействует со средой, выбирая действия на основе политики эпсилон-жадности.
3. Сбор опыта (состояние, действие, вознаграждение, следующее состояние) и сохранение его в буфере воспроизведения опыта.
4. Регулярно берите образцы из буфера воспроизведения для обновления весов сети с использованием метода оптимизации на основе градиентного спуска, обычно оптимизатора Adam.
5. Периодически оценивайте производительность, чтобы оценить сходимость и эффективность изученной политики.

Гиперпараметры, такие как скорость обучения, коэффициент дисконтирования (γ) и размер буфера воспроизведения, были выбраны эмпирически после нескольких симуляций, чтобы обеспечить оптимальную эффективность обучения и стабильную сходимость.

2.6 Реализация симуляции и метрики оценки

MATLAB использовался для реализации среды моделирования перевернутого маятника, обеспечивая точное физическое моделирование и облегчая визуализацию в реальном времени. Модель интегрируется со скриптами MATLAB, отвечающими за реализацию агента обучения с подкреплением, обеспечивая непрерывную связь между симулированной средой и агентом DQN.

Описание MATLAB

MATLAB (сокращение от Matrix Laboratory) — это среда программирования высокого уровня, разработанная MathWorks, широко используемая в инженерии, научных исследованиях и образовании. Она предоставляет интуитивно понятную платформу для численных вычислений, анализа данных, визуализации, разработки алгоритмов и моделирования систем. MATLAB особенно популярен для систем управления, обработки сигналов, машинного обучения и робототехники благодаря своим мощным встроенным инструментам и удобному синтаксису. Одной из его основных сильных сторон является, графическая среда для моделирования и имитации динамических систем с использованием блок-схем. Это позволяет пользователям проектировать сложные системы управления, тестировать алгоритмы и визуализировать поведение системы в реальном времени. В контексте данного исследования MATLAB и используются для создания среды моделирования для перевернутого маятника, реализации и обучения агента обучения с подкреплением и оценки производительности системы в различных условиях. Его гибкость и обширная поддержка библиотек делают MATLAB идеальным инструментом как для теоретических исследований, так и для практической реализации в области проектирования систем управления.

Оценка производительности основана на:

– Длительности эпизода: Длительности, в течение которой агент успешно сохраняет равновесие в течение одного эпизода симуляции.

– Кумулятивное вознаграждение: Совокупные вознаграждения, полученные в ходе эпизодов, отражающие общую эффективность обученной политики.

– Устойчивость: Дисперсия и постоянство углового положения маятника, указывающие на надежность стабилизации.

Эта методологическая основа обеспечивает всестороннее исследование и проверку способности DQN достигать стабильного управления перевернутым маятником в моделируемой среде MATLAB.

3 Результаты

3.1 Обучение модели RL

Модель обучения с подкреплением для стабилизации перевернутого маятника была разработана и обучена с использованием алгоритма Deep Q-Network (DQN) в среде MATLAB. MATLAB предоставляет ряд интерактивных приложений, рисунок 4.1, которые упрощают различные сложные задачи моделирования, имитации и развертывания. Эти приложения позволяют эффективно и интуитивно проводить эксперименты с алгоритмами и моделями без обширного программирования, что значительно сокращает время и усилия по разработке.

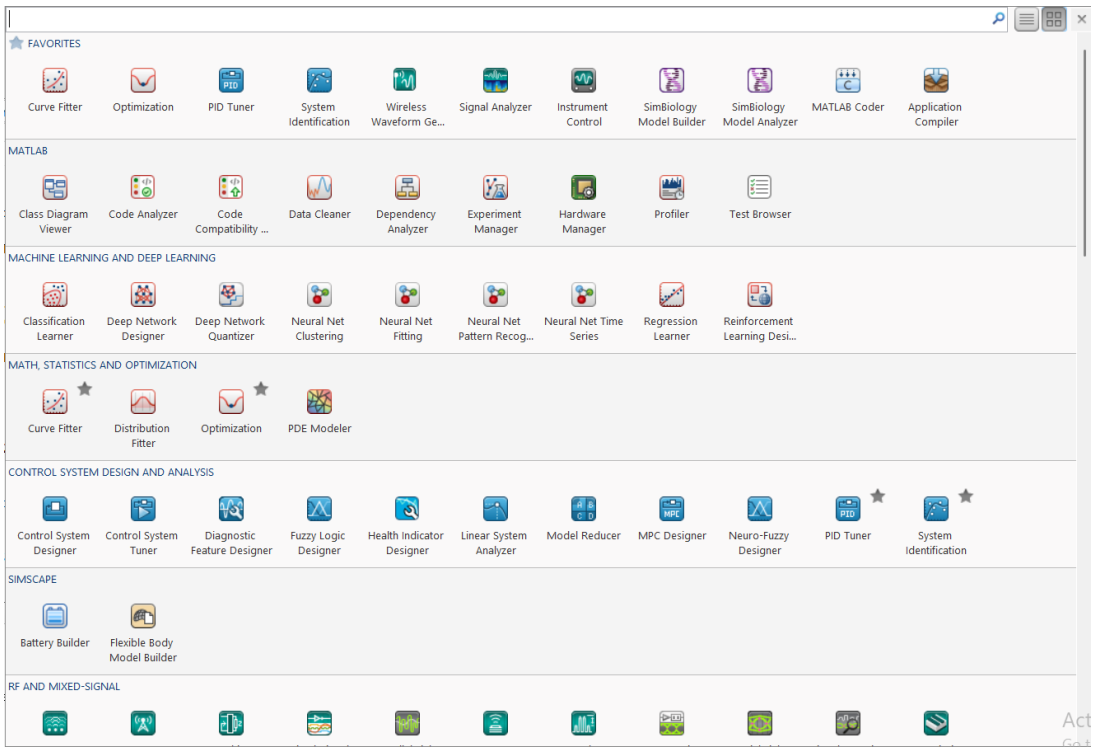


Рисунок 4.1 – Интерактивные приложения Matlab

В частности, для этого проекта использовалось приложение Reinforcement Learning Designer, на рисунке 4.2. Это специализированное приложение позволяет пользователям интерактивно проектировать, обучать, визуализировать и тестировать агентов обучения с подкреплением. Оно объединяет различные

алгоритмы обучения с подкреплением, включая Deep Q-Networks, предоставляя обширные возможности конфигурации и мониторинг процесса обучения в реальном времени. Пользователи получают выгоду от его графического интерфейса, который упрощает определение сред, настройку гиперпараметров и оценку производительности агентов.

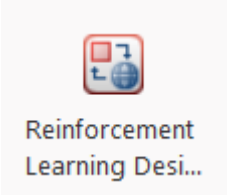


Рисунок 4.2 – Интерактивное приложение Reinforcement Learning Designer

В качестве среды обучения была выбрана среда «Discrete Cart Pole», предоставленная MATLAB. Эта среда имитирует систему перевернутого маятника, состоящую из тележки, которая может двигаться горизонтально, стремясь удерживать маятник в вертикальном положении. Среда предоставляет агенту обучения с подкреплением четыре основных наблюдения, рисунок 4.3:

– Положение тележки (x)
– Скорость тележки (x')
– Угол маятника (θ)
– Угловая скорость маятника (θ')

```
Discrete_CartPole =

Observation Dimension
   CartPole States : [4 1]
Action Dimension
   CartPole Action : [1 1]
```

Рисунок 4.3 – Векторы наблюдения и действии среды Discrete Cart Pole

Агент взаимодействует с окружающей средой, выбирая одно из двух дискретных действий:

– Применить силу -10 Н (переместить тележку влево)

– Приложить силу +10 Н (переместить тележку вправо)

Для процесса обучения агент DQN был настроен со следующими параметрами, на рисунке 4.4:

– Время выборки: 1 секунда

– Коэффициент дисконтирования (γ): 0,99 (высоко ценит будущие награды)

– Среда выполнения: ЦП

– Размер пакета: 64

– Длина буфера опыта: 10 000

– Скорость обучения: 0,0001

– Порог градиента: Бесконечность (без порогового значения)

– Начальный эпсилон: 1 (полное исследование изначально)

– Распад эпсилона: 0,005

– Минимум эпсилона: 0,01 (вероятность конечного исследования)

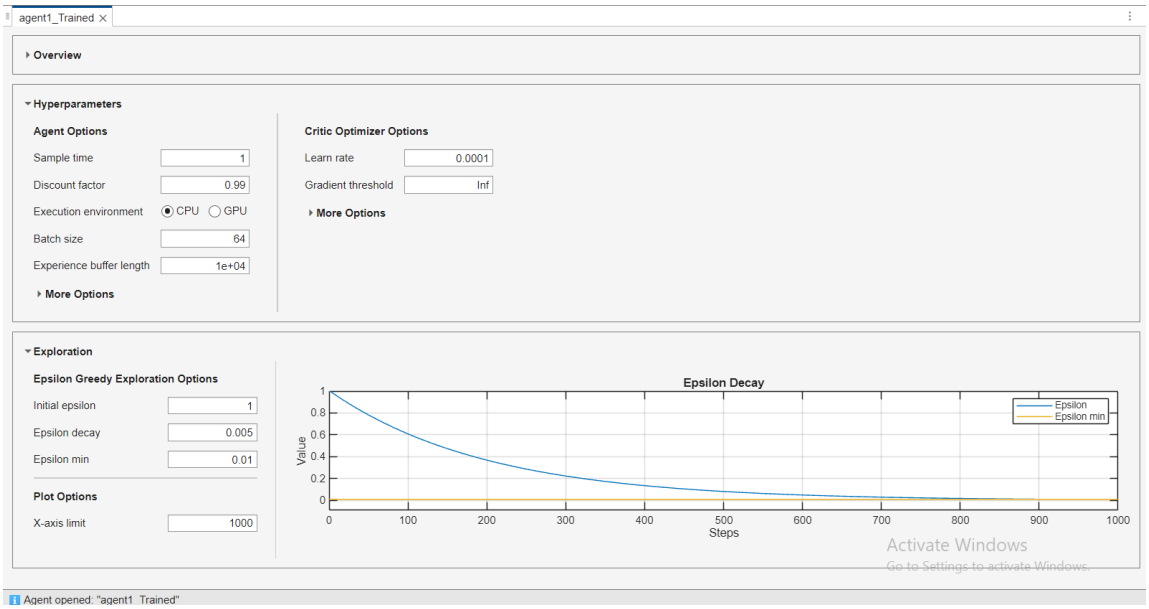


Рисунок 4.4 – Параметры и настройки обучения агента

Обучение продолжалось до тех пор, пока агент не достигал последовательного значения оценки вознаграждения 500, что указывает на успешную стабилизацию перевернутого маятника. На рисунке 4.5 изображен график обучения агента обучения с подкреплением. Этот рубеж был впервые

достигнут после 314 эпизодов, при этом общее время обучения составило 10 минут и 12 секунд.

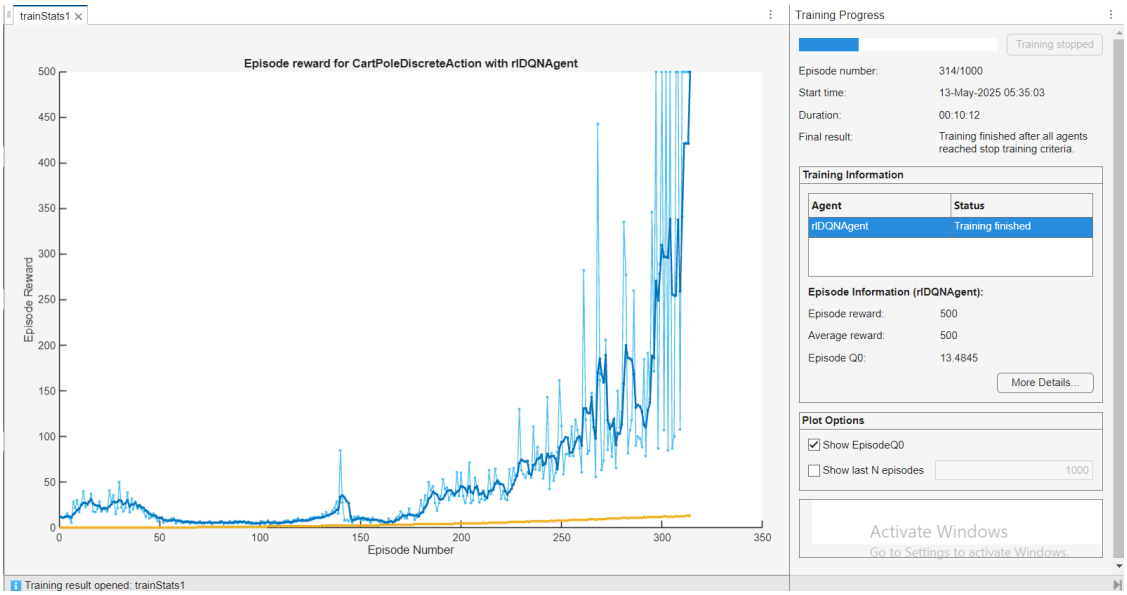


Рисунок 4.5 – График обучения агента обучения с подкреплением

Оценки после обучения проводились с помощью имитационных тестов, подтверждающих эффективность агента. Во время этих имитаций модель обучения с подкреплением успешно поддерживала оценку вознаграждения 500 в 9 из 10 испытаний, рисунок 4.6. Однако в одной имитации эффективность агента была заметно ниже, достигнув оценки приблизительно 100, что подчеркивает случайную нестабильность или неоптимальную конвергенцию политики. Этот результат подчеркивает присущую изменчивость и чувствительность к начальным условиям и стратегиям исследования в обучении с подкреплением.

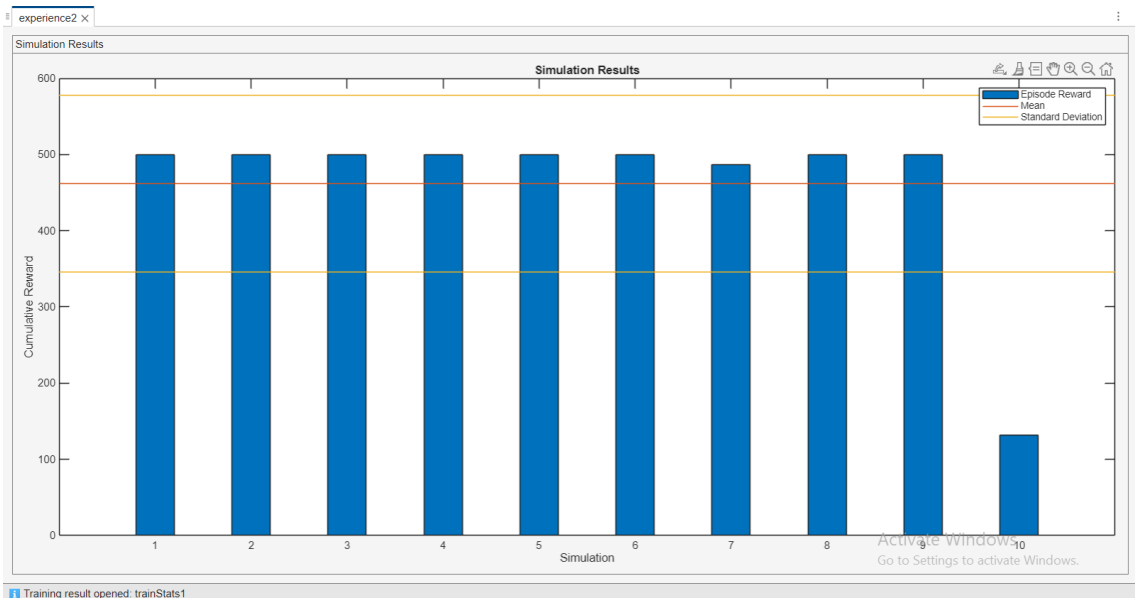


Рисунок 4.6 – Результаты тестов симуляции

3.2 Визуализация

Визуализация производительности обученного агента обучения с подкреплением была выполнена с использованием скрипта MATLAB (ссылка предоставляется отдельно), рисунок 4.7. Перед визуализацией обученный агент должен быть сначала экспортирован в рабочее пространство MATLAB. Моделирование начинается со случайного угла маятника в диапазоне от -0,05 до +0,05 радиан. Визуализация продолжается до тех пор, пока угол маятника не превысит 12 градусов от вертикальной оси или положение тележки не отклонится более чем на 2,4 метра от своего исходного положения, после чего моделирование автоматически прекращается. Рисунок иллюстрирует успешный сценарий стабилизации маятника, в котором агент эффективно поддерживал равновесие. Напротив, рисунок 2 показывает пример моделирования, в котором агент не смог сохранить равновесие, в результате чего маятник превысил порог в 12 градусов, и, таким образом, визуализация прекратилась. На рисунках 4.8, 4.9 можно наблюдать успешные и неуспешные попытки балансировки маятника агентом.

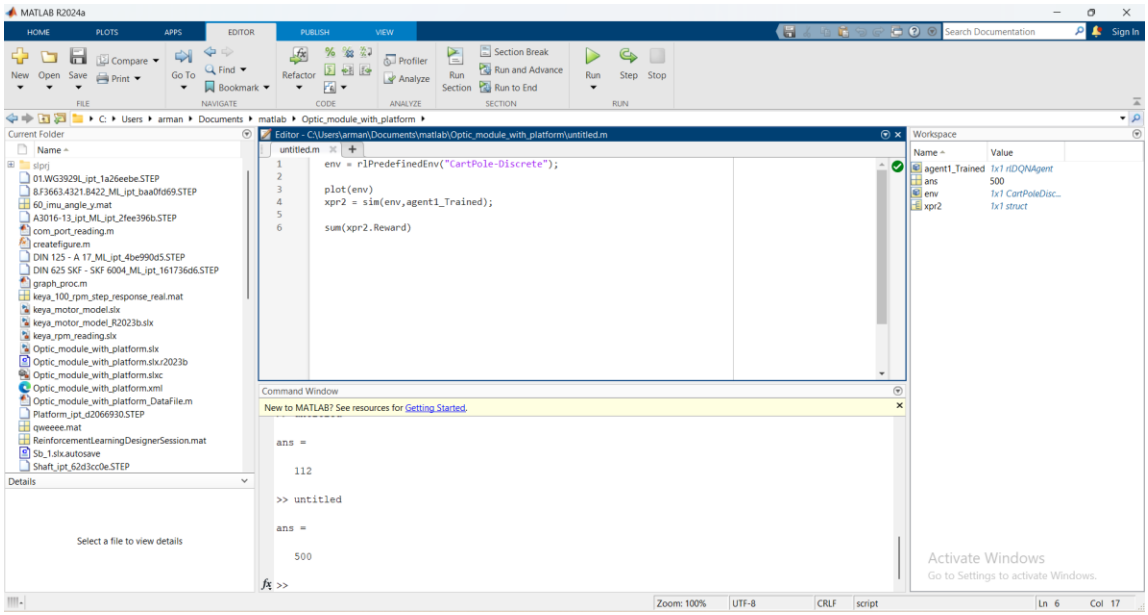


Рисунок 4.7 – Matlab скрипт для выполнения визуализации

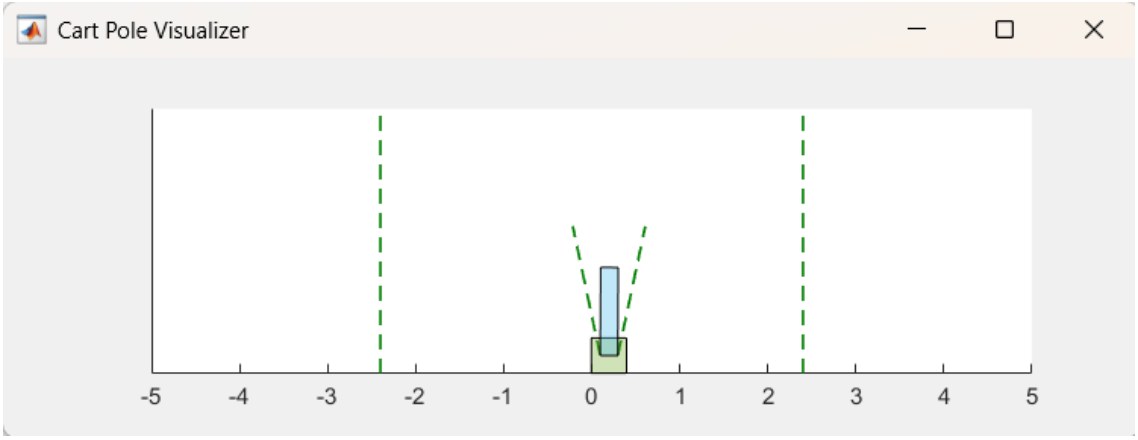


Рисунок 4.8 – Успешная балансировка маятника агентом

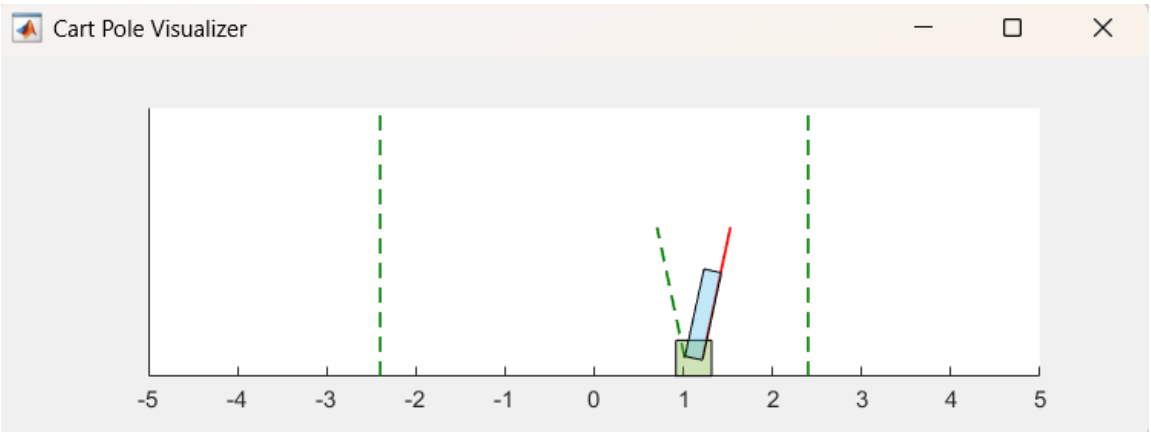


Рисунок 4.9 – Не успешная балансировка маятника агентом

3.3 Анализ производительности модели

Подробный анализ производительности был проведен на основе 7-й симуляции, которая предоставила интересные наблюдения о поведении агента в условиях тестирования. Данные о положении тележки, ее скорости, угле маятника, угловой скорости, дискретных действиях, выполненных моделью, и связанных с ними баллах вознаграждения были тщательно извлечены и изучены.

На рисунке 4.10 проиллюстрирована траектория положения тележки. Было отмечено, что тележка сохраняла отклонение положения в приемлемых пределах, не превышая максимального отклонения в 1 метр и оставаясь в пределах критической границы в 2,4 метра. Это указывает на контролируемое, стабильное горизонтальное движение тележки на протяжении большей части симуляции.

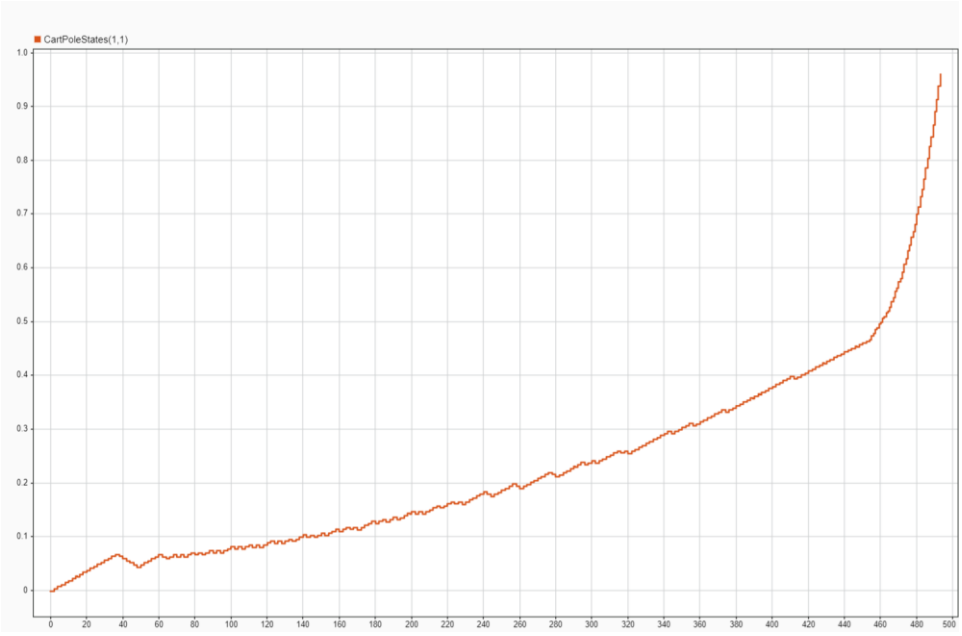


Рисунок 4.10 – График позиции тележки по времени

Скорость тележки изображена на рисунке 4.11, который дает представление о том, как модель регулировала скорость для стабилизации маятника. График скорости демонстрирует плавные и реактивные реакции управления, соответствующие ожиданиям по поддержанию равновесия.

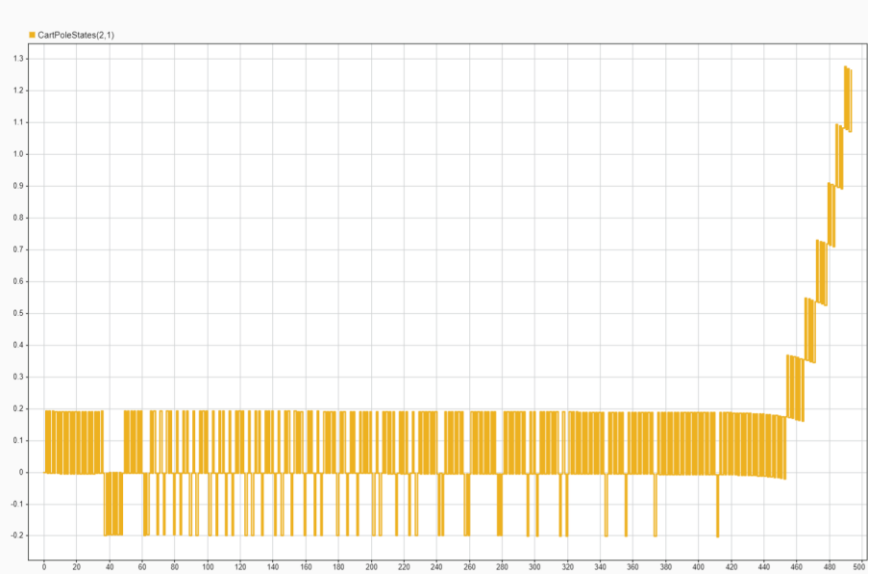


Рисунок 4.11 – График скорости тележки по времени

Угол маятника, представленный в радианах, показан на рисунке 4.12. Первоначально угол маятника оставался в пределах стабильного, небольшого диапазона, отражая эффективный контроль баланса агентом. Однако по мере того, как моделирование продвигалось к более поздним временным шагам, колебания увеличивались, указывая на растущие проблемы в поддержании точного контроля.

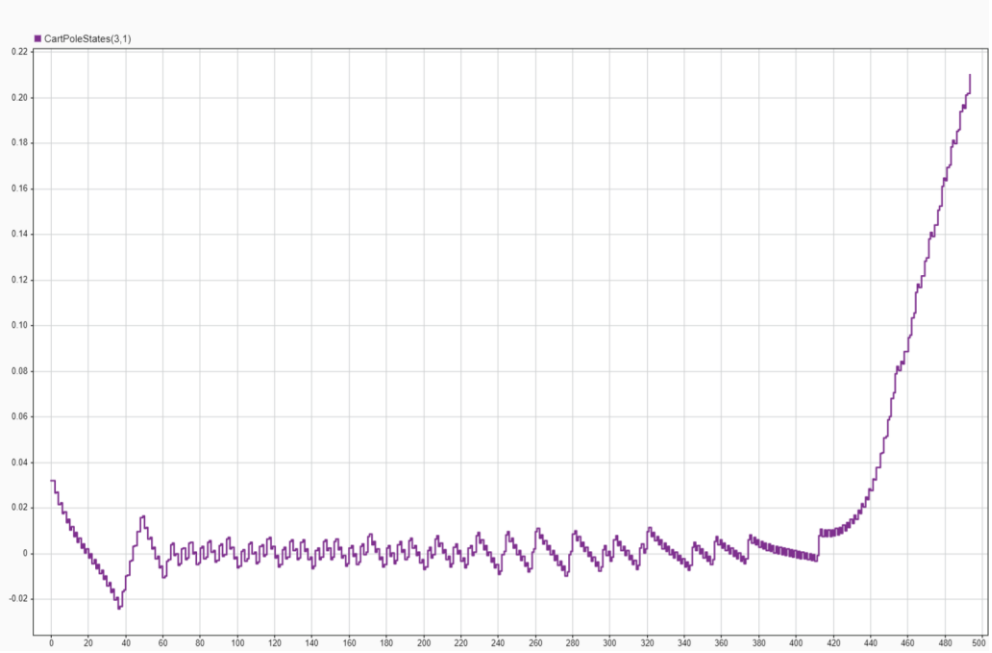


Рисунок 4.12 – График отклонение угла маятника по времени

Угловая скорость маятника, еще один критический фактор, влияющий на устойчивость, изображена на рисунке 4.13. График угловой скорости подтверждает колебания угла, показывая более значительные изменения по мере приближения маятника к неустойчивости.

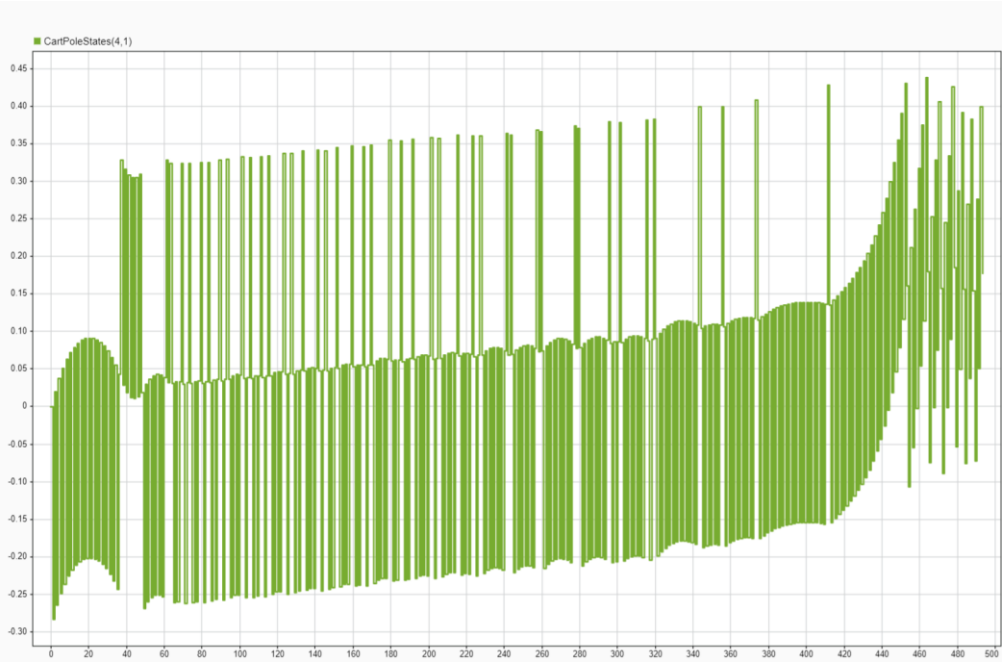


Рисунок 4.13 – График угловой скорости маятника по времени

Рисунок 4.14 иллюстрирует дискретные действия, предпринимаемые моделью, чередуя приложение силы -10 Н и +10 Н. Анализ этих действий помогает понять стратегические решения, принимаемые агентом обучения с подкреплением для противодействия отклонениям в положении и скорости маятника.

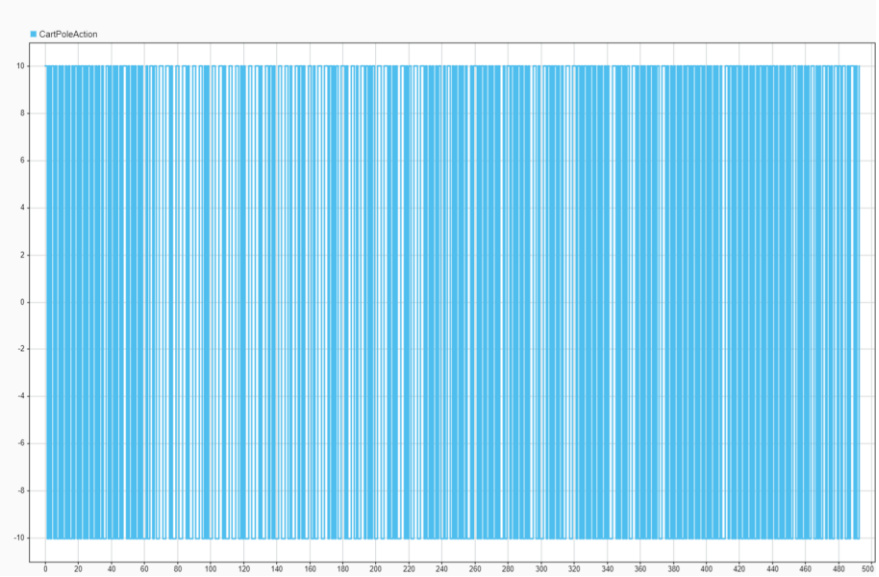


Рисунок 4.14 – График дискретных действии агента по времени

Баллы вознаграждения, полученные агентом на каждом временном шаге, показаны на рисунке 4.15. Вознаграждения демонстрируют успешную стабилизацию примерно до 460-го временного шага, после которого модель не смогла поддерживать требуемую устойчивость, в конечном итоге заставив угол маятника превысить порог в 12 градусов. Эта неудача в конечном итоге привела к неполному максимальному баллу вознаграждения, немного ниже идеального значения 500.

Рисунок 4.15 – График вознаграждении агента по времени

Подробный анализ подчеркивает как сильные стороны, так и ограничения обученной модели, предоставляя важные сведения о процессе принятия решений агентом и условиях, приводящих к успешным и неудачным результатам. Эта тщательная оценка позволяет вносить целенаправленные улучшения в будущие итерации модели.

3.4Результаты

Результатом этой работы стала успешная разработка и обучение агента обучения с подкреплением с использованием приложения MATLAB Reinforcement Learning Designer. Агент был эффективно обучен в контролируемой среде, а его производительность была тщательно протестирована с помощью нескольких симуляций. Визуализации способствовали четкому пониманию поведения агента, иллюстрируя как успешные, так и неудачные попытки балансировки. Анализ производительности данных моделирования предоставил подробные сведения о сильных сторонах и областях для улучшения обученной модели, подтвердив способность модели стабилизировать перевернутый маятник в различных условиях, а также определив условия, ведущие к нестабильности.

3.5Будущие улучшения

Для будущих улучшений изучение альтернативных алгоритмов обучения с подкреплением может значительно повысить эффективность модели. Такие методы, как Proximal Policy Optimization (PPO), Soft Actor-Critic (SAC) или подходы, использующие непрерывные пространства действий, могут предложить потенциально более точные и надежные стратегии управления, особенно в высокодинамичных или шумных средах.

Более того, переход от моделирования к реальным приложениям представляет собой важный следующий шаг. Развертывание обученного агента обучения с подкреплением на физических системах перевернутого маятника

может подтвердить результаты моделирования и обеспечить практическую применимость. Реализация аппаратной интеграции будет включать такие соображения, как точность датчика, задержка, точность привода и устойчивость к шуму. Преодоление разрыва между моделированием и динамикой реального мира имеет важное значение, требуя тщательной калибровки и, возможно, переобучения или тонкой настройки моделей в физическом контексте.

Дальнейшие эксперименты с настройкой гиперпараметров, стратегиями исследования и архитектурами обучения могут повысить адаптивность и обобщение агента в различных сценариях, в конечном итоге достигая превосходной производительности в сложных и непредсказуемых условиях реального мира.

Заключение

Эта работа продемонстрировала эффективность алгоритмов обучения с подкреплением, в частности Deep Q-Network (DQN), в стабилизации перевернутого маятника — эталонной проблемы управления, характеризующейся присущей ей нестабильностью и нелинейной динамикой. Полученные результаты подтвердили, что обучение с подкреплением предлагает многообещающую альтернативу традиционным методам управления на основе моделей, автономно обучаясь оптимальным политикам управления посредством взаимодействия с окружающей средой.

С помощью приложения MATLAB Reinforcement Learning Designer агент обучения с подкреплением был эффективно обучен, визуализирован и тщательно протестирован. Агент достиг высокой производительности, успешно стабилизировав маятник в большинстве симуляций, тем самым подтвердив пригодность метода DQN для приложений управления в реальном времени в динамических и неопределенных средах.

Однако изменчивость производительности, подчеркнутая случайными сбоями, подчеркивает чувствительность моделей обучения с подкреплением к начальным условиям и настройкам параметров. Для устранения этих ограничений были определены будущие усовершенствования, такие как изучение других продвинутых алгоритмов обучения с подкреплением, таких как PPO и SAC, а также перенос изученной модели из симуляций в физические установки перевернутого маятника.

В конечном счете, это исследование продемонстрировало значительный потенциал обучения с подкреплением для улучшения адаптивных и надежных стратегий управления, обеспечивая прочную основу для дальнейших достижений в автономных системах, робототехнике и других сценариях управления высокой сложности.

Ссылки на литературу

1. Steven L. Brunton, J. Nathan Kutz. (б.д.). Data Driven Science & Engineering Machine Learning, Dynamical Systems, and Control. стр. 76.
2. Amin, S. (б.д.). Deep Q-Learning (DQN). *Medium*.
3. Andrzej Jezierski, Jakub Mozaryn & Damian Suski . (2017). A Comparison of LQR and MPC Control Algorithms of an Inverted Pendulum. *Springer Nature*.
4. Arne Traue; Gerrit Book; Wilhelm Kirchgässner; Oliver Wallscheid. (2020). Toward a Reinforcement Learning Environment Toolbox for Intelligent Electric Motor Control. *IEEEXplore*.
5. Ayodeji Okubanjo,Oluwadamilola Oyetola. (2019). DYNAMIC MATHEMATICAL MODELING AND CONTROL ALGORITHMS DESIGN OF AN INVERTED PENDULUM SYSTEM. *DergiPark*.
6. Balancing Pencil Challenge. (б.д.). *Science World*.
7. Bethany Lusch, J. Nathan Kutz & Steven L. Brunton . (2018). Deep learning for universal linear embeddings of nonlinear dynamics. *Nature*.
8. Chengrun Qiu; Yang Hu; Yan Chen; Bing Zeng. (2019). Deep Deterministic Policy Gradient (DDPG)-Based Energy Harvesting Wireless Communications. *IEEEXplore*.
9. Control Systems - Introduction. (б.д.). *Tutorialspoint*.
10. Elisa Sara Varghese, Anju K Vincent and V Bagyaveereswaran. (2017). Optimal control of inverted pendulum system using PID controller, LQR and MPC. *IOPScience*.
11. Guirec Maloisel, Christian Schumacher, Espen Knoop. (2023). Optimal Design of Robotic Character Kinematics. *Disney Research*.
12. Han Shen; Kaiqing Zhang; Mingyi Hong; Tianyi Chen. (2023). Towards Understanding Asynchronous Advantage Actor-Critic: Convergence and Linear Speedup. *IEEEXplore*.
13. Hazem, Z. B. (2024). Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system. *Springer*.
14. Hyun-Kyo Lim; Ju-Bong Kim; Chan-Myung Kim; Gyu-Young Hwang; Ho-bin Choi; Youn-Hee Han. (2020). Federated Reinforcement Learning for Controlling Multiple Rotary Inverted Pendulums in Edge Computing Environments. *IEEEXplore*.

15. Ian Osband, Charles Blundell, Alexander Pritzel, Benjamin Van Roy. (2016). Deep Exploration via Bootstrapped DQN. *Neurips*, стр. 9.

16. Ivan Sorokin, Alexey Seleznev, Mikhail Pavlov, Aleksandr Fedorov, Anastasiia Ignateva. (2015). Deep Attention Recurrent Q-Network. *Arxiv*.

17. Jianqing Fan, Zhaoran Wang, Yuchen Xie, Zhuoran Yang. (2020). A Theoretical Analysis of Deep Q-Learning. *MLR*.

18. Job van Amerongen, Peter Breedveld. (2003). Modelling of physical systems for the design and control of mechatronic systems. *Elsevier*.

19. Li, Y. (2017). Deep Reinforcement Learning: An Overview. *Arxiv*.

20. Melrose Roderick, James MacGlashan, Stefanie Tellex. (2017). Implementing the Deep Q-Network. *Arxiv*.

21. Michael Hesse, Julia Timmermann, Eyke Hullermeier, Ansgar Trachtlet. (2018). A reinforcement Learning Strategy for the Swing-Up of the Double Pendulum on a Cart. *Elsevier*, стр. 6.

22. Mohit Sewak. (2019). Actor-Critic Models and the A3C. *Springer*.

23. Olfa Boubaker. (2013). The Inverted Pendulum Benchmark in Nonlinear Control Theory: A Survey. *Sage Journals*.

24. Přemysl Strakoš; Jiří Tůma. (2017). Mathematical modelling and controller design of inverted pendulum. *IEEEXplore*.

25. R, R. (2021). Design of State Feedback LQR Based Dual Mode Fractional Order PID Controller Using Inertia Weighted PSO Algorithm with Reference Model: For Control of an Underactuated System With and Without Disturbance.

26. Ruben Grandia, Espen Knoop, Michael A. Hopkins, Georg Wiedebach, Jared Bishop, Steven Pickles, David M¨ uller, and Moritz B¨ acher. (2024). Design and Control of a Bipedal Robotic Character. *Disney Research, Switzerland, Disney Research, USA, Walt Disney Imagineering R&D, USA*, стр. 14.

27. Ruben Hernandez, Ramon Garcia-Hernandez, Francisco Jurado. (2024). Modeling, Simulation, and Control of a Rotary Inverted Pendulum: A Reinforcement Learning-Based Control Approach. *MDPI*, стр. 29.

28. Sardor Israilov, Li Fu, Jesus Sanchez-Rodriguez, Franco Fusco, Guillaume Allibert, Christophe Raufaste, Mederic Argentina. (2023). Reinforcement learning approach to control an inverted pendulum: A general framework for educational purposes. *PLOS*, стр. 15.

29. Sebastian Höfer; Kostas Bekris; Ankur Handa; Juan Camilo Gamboa; Melissa Mozifian; Florian Golemo. (2021). Sim2Real in Robotics and Automation: Applications and Challenges. *IEEEXplore*.

30. Taegun Lee, Doyoon Ju, Young Sam Lee. (2025). Transition Control of a Double-Inverted Pendulum System Using Sim2Real Reinforcement Learning. *MDPI*, стр. 21.

31. Team, A. E. (б.д.). Reinforcement Learning Explained: Overview, Comparisons and Applications in Business. *AltexSoft*.

32. Thomas L. Mohren; Thomas L. Daniel; Steven L. Brunton. (2020). Learning Precisely Timed Feedforward Control of the Sensor-Denied Inverted Pendulum. *Arxiv*.

33. Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, Sergey Levine. (2018). Soft Actor-Critic Algorithms and Applications. *Arxiv*.

34. Volodymyr Mnih, Koray Kavukcuoglu, David Silver. (2015). Human-level control through deep reinforcement learning. *Nature portfolio*.

35. Yuhui Wang, Hao He, Xiaoyang Tan. (2020). Truly Proximal Policy Optimization. *MLR*, стр. 10.

Приложение А

```
env = rlPredefinedEnv("CartPole-Discrete");

plot(env)
xpr2 = sim(env, agent1_Trained);

sum(xpr2.Reward)
```

Matlab скрипт для запуска визуализации

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН НЕКОММЕРЧЕСКОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО «КАЗАХСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ К.И. САТПАЕВА»

ОТЗЫВ

Дипломной работы

Студента по специальности 6B07111 – «Робототехника и Мехатроника»

Куанаев Арманжан Ахатұлы

На тему: «Обратный маятник»

Дипломная работа посвящена исследованию одной из классических задач теории управления — стабилизации обратного маятника — с использованием современных методов обучения с подкреплением (Reinforcement Learning), в частности, алгоритма Deep Q-Network (DQN).

Целью работы является демонстрация возможности замены традиционных методов управления на подход, основанный на данных и взаимодействии с окружающей средой. Работа носит актуальный и научно-практический характер, так как методы RL находят всё большее применение в задачах управления реальными объектами в условиях неопределённости и динамики.

В ходе выполнения работы студент продемонстрировал владение теоретическими знаниями в области системах управления и искусственного интеллекта. С помощью среды MATLAB Reinforcement Learning Designer был реализован и обучен агент, успешно стабилизирующий систему в большинстве симуляций. Проведённая визуализация, тестирование и анализ подтверждают эффективность и применимость алгоритма DQN для задач стабилизации.

Работа структурирована логично, изложена грамотно и соответствует требованиям, предъявляемым к выпускным квалификационным работам. Все поставленные задачи выполнены в полном объёме.

Выпускная квалификационная работа соответствует требованиям, предъявляемым к ВКР бакалавра, и заслуживает оценки «отлично». Считаю, что студент заслуживает присуждения академической степени бакалавра техники и технологии по направлению подготовки 6B07111 — Робототехника и мехатроника.

Научный руководитель:
Старший преподаватель
Рахметова П. М.
«6» июня 2025 г.

Отчет подобия

Метаданные

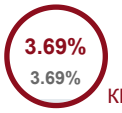
Название организации
Satbayev University
Название
Обратный Маятник

Автор
Куанаев Арманжан Ахатұлы
Научный руководитель / Эксперт
Перизат Рахметова

Подразделение
ИАиИТ

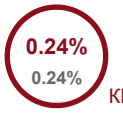
Объем найденных подобий

КП-ия определяют, какой процент текста по отношению к общему объему текста был найден в различных источниках.. Обратите внимание!Высокие значения коэффициентов не означают плагиат. Отчет должен быть проанализирован экспертом.



КП1

КП2



КЦ

25
Длина фразы для коэффициента подобия 2

5471
Количество слов

44011
Количество символов

Тревога

В этом разделе вы найдете информацию, касающуюся текстовых искажений. Эти искажения в тексте могут говорить о ВОЗМОЖНЫХ манипуляциях в тексте. Искажения в тексте могут носить преднамеренный характер, но чаще, характер технических ошибок при конвертации документа и его сохранении, поэтому мы рекомендуем вам подходить к анализу этого модуля со всей долей ответственности. В случае возникновения вопросов, просим обращаться в нашу службу поддержки.

Замена букв		1
Интервалы		0
Микропробелы		1
Белые знаки		0
Парафразы (SmartMarks)	a	8

Подобия по списку источников

Ниже представлен список источников. В этом списке представлены источники из различных баз данных. Цвет текста означает в каком источнике он был найден. Эти источники и значения Коэффициента Подобия не отражают прямого плагиата. Необходимо открыть каждый источник и проанализировать сожержание и правильность оформления источника.

10 самых длинных фраз — Цвет текста

ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ И АДРЕС ИСТОЧНИКА URL (НАЗВАНИЕ БАЗЫ)	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)
1	https://jesit.springeropen.com/articles/10.1186/s43067-022-00050-5	19 0.35 %
2	Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system Ben Hazem, Zied;	19 0.35 %
3	https://forum.robotis.com/t/walt-disney-imagineering-design-and-control-of-a-bipedal-robotic-character/6036	16 0.29 %

4	https://exascale.info/assets/pdf/students/2022_MSc_Nicolas_Roguet.pdf	16 0.29 %
5	Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system Ben Hazem, Zied;	16 0.29 %
6	Glusco_-_master_thesis [redacted] 7/9/2024 Ukrainian Catholic University (Ukrainian Catholic University)	13 0.24 %
7	https://forum.robotis.com/t/walt-disney-imagineering-design-and-control-of-a-bipedal-robotic-character/6036	12 0.22 %
8	https://codelabsacademy.com/ru/blog/understanding-supervised-and-unsupervised-learning	11 0.20 %
9	https://arxiv.org/pdf/2403.02290	10 0.18 %
10	https://www.altexsoft.com/blog/reinforcement-learning-explained-overview-comparisons-and-applications-in-business/	10 0.18 %

из базы данных RefBooks (0.64 %)

ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)
Источник: Paperity		
1	Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system Ben Hazem, Zied;	35 (2) 0.64 %

из домашней базы данных (0.00 %)

ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)

из программы обмена базами данных (0.48 %)

ПОРЯДКОВЫЙ НОМЕР	НАЗВАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)
1	Glusco_-_master_thesis [redacted] 7/9/2024 Ukrainian Catholic University (Ukrainian Catholic University)	13 (1) 0.24 %
2	Разработка сайта по продаже мобильных телефонов [redacted] 3/31/2025 University Miras (University Miras)	8 (1) 0.15 %
3	Optimale batterij opslag gebruik makend van AI [redacted] 6/17/2024 Universiteit Gent (Plato)	5 (1) 0.09 %

из интернета (2.58 %)

ПОРЯДКОВЫЙ НОМЕР	ИСТОЧНИК URL	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)
1	https://forum.robotis.com/t/walt-disney-imagineering-design-and-control-of-a-bipedal-robotic-character/6036	41 (4) 0.75 %
2	https://arxiv.org/pdf/2403.02290	30 (4) 0.55 %
3	https://jesit.springeropen.com/articles/10.1186/s43067-022-00050-5	19 (1) 0.35 %
4	https://exascale.info/assets/pdf/students/2022_MSc_Nicolas_Roguet.pdf	16 (1) 0.29 %

5	https://flibusta.su/book/348469-neyroseti-osnovyi/read/	14 (2) 0.26 %
6	https://codelabsacademy.com/ru/blog/understanding-supervised-and-unsupervised-learning	11 (1) 0.20 %
7	https://www.altexsoft.com/blog/reinforcement-learning-explained-overview-comparisons-and-applications-in-business/	10 (1) 0.18 %

Список принятых фрагментов (нет принятых фрагментов)

ПОРЯДКОВЫЙ НОМЕР	СОДЕРЖАНИЕ	КОЛИЧЕСТВО ИДЕНТИЧНЫХ СЛОВ (ФРАГМЕНТОВ)