

Санкт-Петербургский государственный университет

Кафедра прикладной кибернетики

Группа 24.Б82-мм

# Алгоритмы тікмік

***Черняков Евгений Олегович***

Отчёт по учебной практике  
в форме «Сравнение»

Научный руководитель:  
профессор, научный сотрудник кафедры прикладной кибернетики Мокаев Р. Н.

Санкт-Петербург  
2025

# Оглавление

<b>Введение</b>	<b>3</b>
<b>1. Постановка задачи</b>	<b>5</b>
<b>2. Обзор</b>	<b>6</b>
2.1. Обзор существующих решений . . . . .	6
2.2. Обзор используемых технологий . . . . .	9
2.3. Выводы . . . . .	11
<b>3. Описание решения</b>	<b>13</b>
3.1. Первая задача . . . . .	13
3.2. Вторая задача . . . . .	20
3.3. Третья задача . . . . .	23
<b>4. Эксперимент</b>	<b>35</b>
4.1. Условия эксперимента . . . . .	35
4.2. Исследовательские вопросы . . . . .	35
4.3. Метрики . . . . .	36
4.4. Результаты . . . . .	36
4.5. Обсуждение результатов . . . . .	38
4.6. Угрозы нарушения корректности (опциональный) . . . . .	38
4.7. Воспроизводимость эксперимента . . . . .	38
<b>Список литературы</b>	<b>40</b>

# Введение

Задача стабилизации перевернутого маятника на подвижной тележке (Cart-Pole) является классическим эталоном в теории автоматического управления и обучении с подкреплением. Она моделирует неустойчивые нелинейные системы, принципы управления которыми лежат в основе работы балансирующих роботов и ракет-носителей на старте. Традиционные методы решения, такие как ПИД-регуляторы или LQR-алгоритмы, требуют точного знания математической модели объекта. Однако с развитием искусственного интеллекта всё большую популярность приобретают методы, способные обучаться управлению через взаимодействие со средой, не требуя явного задания уравнений динамики. Существующие исследования широко освещают применение глубокого обучения с подкреплением (Deep RL) для этой задачи, демонстрируя возможности сложных нейросетевых архитектур.

В то же время, в учебной и технической литературе часто наблюдается пробел в сравнительном анализе эффективности базовых стохастических методов оптимизации по сравнению с градиентными методами начального уровня. Нередко сложные алгоритмы применяются там, где достаточно простейшего случайного поиска. Остается открытym вопросом: оправдано ли усложнение алгоритма (переход от случайногo поиска к градиентной политике) для системы с малым количеством степеней свободы, и как именно соотносятся затраты на вычисления и качество стабилизации при реализации физической модели «с нуля», без использования готовых библиотек симуляции.

Целью данной работы является выявление сравнительной эффективности алгоритмов различной природы — базового случайного поиска (Basic Random Search), метода восхождения на холм (Hill Climbing) и градиентной политики (Policy Gradient) — в задаче балансировки стержня. Необходимо определить, какой из подходов обеспечивает наилучший баланс между скоростью обучения и устойчивостью удержания маятника. Работа призвана продемонстрировать, что для определенных классов задач простые безградиентные методы могут составлять

конкуренцию более сложным подходам RL, а также показать влияние точности физической симуляции на процесс обучения агентов.

Для реализации поставленной задачи в работе разрабатывается программная модель динамической системы, основанная на численном решении дифференциальных уравнений движения стержня и тележки. В данную среду интегрируются и настраиваются алгоритмы Basic Random Search, Hill Climbing и Policy Gradient. В ходе экспериментов проводится обучение агентов с фиксацией ключевых метрик: времени сходимости алгоритма и длительности удержания стержня в вертикальном положении. На основе полученных данных будет проведен сравнительный анализ и сформулированы выводы о применимости каждого из методов для задач стабилизации неустойчивых динамических систем.

# **1. Постановка задачи**

Целью работы является проведение сравнительного анализа эффективности алгоритмов стохастической оптимизации и методов обучения с подкреплением при решении задачи стабилизации стержня на подвижном основании. Для её выполнения были поставлены следующие задачи:

1. математически описать динамику системы «тележка-стержень», составив дифференциальные уравнения движения;
2. создать программную среду, эмулирующую поведение системы во времени на основе разработанной математической модели, без использования сторонних физических движков;
3. программно реализовать алгоритмы: Basic Random Search, Hill Climbing, Policy Gradient;
4. провести серию экспериментов по обучению агентов каждым из алгоритмов, зафиксировав метрики скорости сходимости и длительности удержания стержня в вертикальном положении;
5. сравнить полученные данные, выявить преимущества и недостатки каждого алгоритма и оценить целесообразность использования градиентных методов для задач данной размерности.

## 2. Обзор

Целью данного раздела является формирование необходимого теоретического фундамента для понимания работы и обоснование методологического выбора, сделанного в рамках исследования. Обзор систематизирует известную информацию о механической системе «тележка-стержень», существующих подходах к ее стабилизации и принципах работы выбранных алгоритмов обучения.

Обзор был выполнен с использованием следующих методов:

- Систематический поиск: Использовались научные базы данных (например, Google Scholar, Scopus, ResearchGate) по ключевым запросам, охватывающим как классическую теорию управления, так и современное обучение с подкреплением (например, «Cart-Pole stabilization», «Inverted Pendulum control», «Policy Gradient vs Random Search»).
- Оценка релевантности: В фокусе внимания находились статьи и монографии, описывающие математическую модель системы с распределенной массой (стержень), а также работы, напрямую сравнивающие эффективность стохастической оптимизации (Random Search, Hill Climbing) с градиентными методами.
- Идентификация технологий: Изучались принципы работы выбранных алгоритмов (PG, HC, BRS) и современные методы их реализации на языке Python.

### 2.1. Обзор существующих решений

Обзор существующих решений направлен на систематизацию научного и инженерного опыта, накопленного в области стабилизации неустойчивых динамических систем, и, в частности, механической системы «тележка-стержень». Это позволяет избежать дублирования уже известных подходов и четко определить место данной работы среди предыдущих исследований.

### **2.1.1. Математическая модель «Тележка-Стержень»**

Система «тележка-стержень» является классическим объектом исследования в теории автоматического управления (ТАУ) и представляет собой идеальный пример нелинейной, существенно неустойчивой системы. Математическое описание основывается на законах Ньютона или уравнениях Лагранжа. В большинстве ранних работ используется модель, где масса стержня считается точечной, однако для точной симуляции требуется модель с учетом распределенной массы (стержня), что усложняет расчеты, но повышает точность реализации.

### **2.1.2. Классические подходы к стабилизации**

Исторически задача решалась с помощью методов, основанных на глубоком знании математической модели:

- 1. Линеаризация и LQR-регулирование (Linear Quadratic Regulator).** Этот подход является золотым стандартом в ТАУ. Преимущество LQR заключается в том, что он гарантирует оптимальность управления (минимизацию заданного квадратичного функционала) для линеаризованной системы, обеспечивая высокую скорость реакции вблизи точки равновесия. Недостаток метода в том, что он требует точной линеаризации уравнений вокруг вертикального положения и может стать неэффективным при больших начальных отклонениях стержня.
- 2. ПИД-регулирование (PID Control).** Преимуществом ПИД-регулятора является его простота и универсальность. Он не требует полного знания уравнений движения. Недостаток заключается в сложности подбора трех коэффициентов ( $K_p, K_i, K_d$ ) для достижения устойчивости в широком диапазоне состояний, особенно в нелинейных зонах.
- 3. Управление на основе фазового пространства.** Некоторые решения используют переключение между различными линейными стратегиями в зависимости от текущего состояния (угла и уг-

ловой скорости), но их разработка трудоемка и слабо масштабируется на более сложные робототехнические системы.

### **2.1.3. Современные подходы на основе обучения с подкреплением (RL)**

С развитием вычислительных мощностей и нейронных сетей методы RL стали доминирующими в задачах, требующих адаптивного управления.

1. **Value-Based методы (DQN, Q-learning).** Эти алгоритмы эффективны для дискретного пространства действий, что применимо к задаче, если управляющую силу  $F(t)$  дискретизировать (например,  $F \in \{-10, 0, 10\}$ ). Однако их сложно масштабировать, и они требуют аппроксимации функции ценности, что часто приводит к нестабильности обучения.
2. **Policy-Based методы (Policy Gradient, Actor-Critic, PPO).** Эти методы непосредственно оптимизируют политику  $\pi$ , что делает их идеальными для задач с непрерывным пространством действий и состояний. Преимущество Policy Gradient (PG), используемого в данной работе, заключается в его фундаментальности, позволяя агенту обучаться, лишь наблюдая за результатами своих действий (наградой), без явного моделирования среды. Недостаток PG — высокая дисперсия градиента, требующая большого числа выборок для стабилизации.

### **2.1.4. Обоснование выбора методологии**

Данное исследование не ставит целью изобретение нового алгоритма стабилизации. Вместо этого, работа фокусируется на методологическом сравнении трех фундаментально разных классов оптимизации — наивного поиска (BRS), локального поиска (HC) и градиентной оптимизации (PG). В существующей литературе, как правило, сравниваются только передовые алгоритмы RL (A2C vs. PPO [1]), тогда как прямое

сравнение их эффективности с простыми стохастическими методами на базовой, но точно реализованной физической модели встречается редко. Этот анализ позволит определить, насколько сложность PG оправдана по сравнению с его простейшими аналогами для систем низкой размерности.

## 2.2. Обзор используемых технологий

Данный подраздел посвящен обзору и обоснованию выбора тех технологических решений и алгоритмов, которые будут использованы для программной реализации задачи стабилизации стержня. Цель — показать, почему выбранные алгоритмы (BRS, HC, PG) являются наилучшим набором для достижения целей исследования, а выбранные инструменты — оптимальными для реализации.

### 2.2.1. Сравнение и выбор алгоритмов оптимизации

В рамках работы проводится сравнительный анализ трех методов, представляющих разные уровни сложности и подходы к оптимизации функций. Обзорная литература [3, 4] показывает, что эти методы формируют логическую иерархию, позволяющую провести глубокое методологическое сравнение:

1. **Basic Random Search (BRS):** Простейший безградиентный метод.
2. **Hill Climbing (HC):** Эвристический метод, использующий локальный случайный поиск для ускорения сходимости BRS.
3. **Policy Gradient (PG):** Градиентный метод обучения с подкреплением, который теоретически должен превосходить безградиентные методы по скорости сходимости в задачах оптимизации [4].

Для оценки алгоритмов использовались следующие критерии, релевантные целям работы: вычислительная сложность, скорость сходимости и риск застревания в локальном оптимуме.

Алгоритм	Класс	Вычисл. сложность	Риск локального оптимума
Basic Random Search	Безградиентный	Низкая	Низкий
Hill Climbing	Безградиентный	Низкая	Высокий
Policy Gradient	Градиентный (RL)	Средняя	Средний

Таблица 1: Сравнение алгоритмов оптимизации

**Обоснование выбора:** Данный набор выбран потому, что он позволяет ответить на ключевой вопрос исследования: оправдывает ли рост вычислительной сложности (переход от НС к PG) пропорциональный рост скорости сходимости в задаче стабилизации низкой размерности. BRS и НС выступают в роли надежных, но простых контрольных групп для оценки эффективности градиентного подхода.

### 2.2.2. Выбор платформы реализации

Для реализации физической модели и алгоритмов необходимо выбрать подходящую программную среду и язык. Для работ в области машинного обучения стандартным выбором является язык **Python**, несмотря на то, что **C/C++** обеспечивает более высокую производительность для чистого численного интегрирования дифференциальных уравнений [2].

Платформа	Производительность	Удобство ML/RL	Скорость разработки
C/C++	Высокая	Низкое	Низкая
Python	Средняя	Высокое	Высокая

Таблица 2: Сравнение платформ для реализации

**Обоснование выбора:** Для задачи стабилизации стержня, которая является системой низкой размерности, высокая производительность C/C++ не является критичной. В то же время, основная цель работы — сравнение алгоритмов обучения. Использование языка **Python**

в связи с фреймворком **PyTorch** значительно ускоряет разработку, упрощает реализацию Policy Gradient за счет автоматического дифференцирования и обеспечивает лучшую читаемость кода, что критически важно для курсовой работы.

### 2.2.3. Используемые библиотеки и инструменты

Для программной реализации среды и алгоритмов будет использован следующий стек технологий:

- **Язык программирования: Python.**
- **Численные расчеты и симуляция:** Библиотека **NumPy**. Используется для эффективного хранения и оперирования матрицами весовых коэффициентов агентов.
- **Реализация алгоритма Policy Gradient:** Фреймворк **PyTorch**. Выбран для удобства работы со стохастическими распределениями (сэмплирование действий агента), необходимых для алгоритма Policy Gradient.
- **Визуализация и анализ данных:** Библиотека **Matplotlib**. Будет применяться для визуализации экспериментов, построения графиков сходимости, визуализации метрик и сравнения результатов экспериментов.

## 2.3. Выводы

Обзор существующих решений (раздел 2.1) показал, что, хотя задача стабилизации стержня решена классическими и современными методами, отсутствует прямое сравнение выбранных алгоритмов (BRS, HC, PG) на самописной физической модели. Обзор технологий (раздел 2.2) подтвердил, что язык **Python** и его библиотеки оптимальны для реализации, поскольку приоритет отдается удобству и скорости реализации сложных алгоритмов машинного обучения перед микрооптимизацией

физической симуляции. Таким образом, теоретическая база для выполнения поставленных задач полностью сформирована.

### 3. Описание решения

В данной главе описывается последовательный процесс реализации системы автоматической стабилизации стержня. Решение задачи декомпозировано на три взаимосвязанных этапа, отражающих переход от теоретической модели к исполняемому программному коду.

В рамках **первого этапа** (подраздел 3.1) выполняется математическое моделирование. На основе законов классической механики выводятся нелинейные дифференциальные уравнения движения, которые ложатся в основу физической симуляции.

На **втором этапе** (подраздел 3.2) разрабатывается программная среда. На базе выведенных уравнений реализован класс `Model`, который выполняет численное интегрирование динамики системы методом Рунге-Кутты. Данный модуль выступает в роли /quoteСреды (Environment) в терминологии обучения с подкреплением.

**Третий этап** (подраздел 3.3) посвящен алгоритмической реализации агентов управления. Были запрограммированы и интегрированы в среду три метода различной природы:

- **Basic Random Search (BRS)** — в качестве базового метода глобальной стохастической оптимизации.
- **Hill Climbing (HC)** — как улучшенная эвристика локального поиска.
- **Actor-Critic (AC)** — как представитель градиентных методов RL, способный обучаться online.

Весь программный комплекс реализован на языке **Python** с активным использованием библиотеки **NumPy** для ускорения векторных вычислений.

#### 3.1. Первая задача

Для описания динамики системы «перевернутый маятник на тележке» воспользуемся вторым законом Ньютона. Система состоит из те-

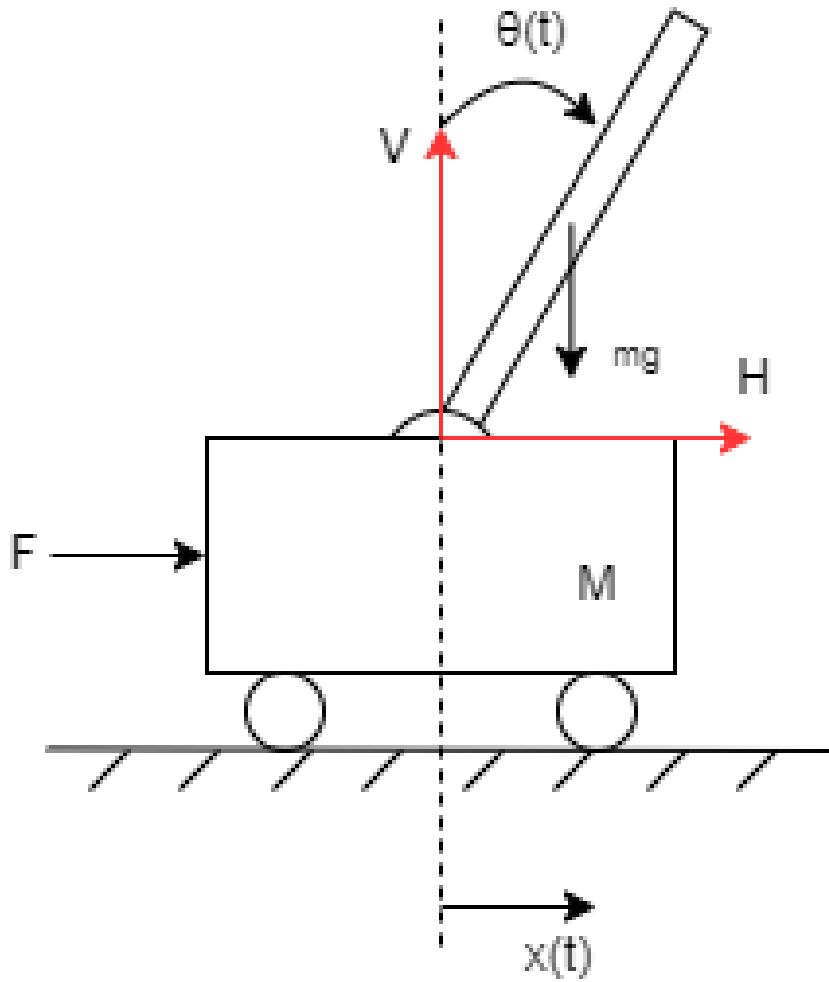


Рис. 1: Схема системы «Перевернутый маятник на тележке» с действующими силами.

лежки массой  $M$ , движущейся горизонтально, и стержня массой  $m$  и длиной  $l$ , закрепленного на тележке через шарнир.

Введем следующие обозначения:

- $x(t)$  — горизонтальная координата тележки;
- $\theta(t)$  — угол отклонения стержня от вертикали (по часовой стрелке);
- $F$  — управляющая сила, приложенная к тележке;
- $H, V$  — горизонтальная и вертикальная компоненты силы реакции в шарнире;

- $a = l/2$  — расстояние от оси шарнира до центра масс стержня;
- $I$  — момент инерции стержня относительно точки крепления.

### 3.1.1. Кинематика системы

Координаты центра масс стержня  $(x_c, y_c)$  выражаются через обобщенные координаты системы  $x$  и  $\theta$ :

$$x_c = x + a \sin \theta, \quad y_c = a \cos \theta. \quad (1)$$

Дважды продифференцировав эти выражения по времени, получим проекции ускорения центра масс стержня:

$$\begin{cases} \ddot{x}_c = \ddot{x} + a \cos \theta \ddot{\theta} - a \sin \theta \dot{\theta}^2, \\ \ddot{y}_c = -a \sin \theta \ddot{\theta} - a \cos \theta \dot{\theta}^2. \end{cases} \quad (2)$$

### 3.1.2. Уравнения динамики

Рассмотрим силы, действующие на тела системы.

**1. Движение тележки.** На тележку действуют управляющая сила  $F$  и горизонтальная реакция стержня  $H$  (по третьему закону Ньютона, направленная противоположно силе, действующей на стержень):

$$M\ddot{x} = F - H. \quad (3)$$

**2. Движение стержня.** Запишем уравнения движения центра масс стержня в проекциях на оси координат:

$$\begin{cases} m\ddot{x}_c = H, \\ m\ddot{y}_c = V - mg. \end{cases} \quad (4)$$

Подставим выражения для ускорений (2) в систему (4):

$$\begin{cases} m(\ddot{x} + a \cos \theta \ddot{\theta} - a \sin \theta \dot{\theta}^2) = H, \\ m(-a \sin \theta \ddot{\theta} - a \cos \theta \dot{\theta}^2) = V - mg. \end{cases} \quad (5)$$

Исключим силу реакции  $H$ . Сложим уравнение тележки (3) и первое уравнение стержня:

$$M\ddot{x} + m(\ddot{x} + a \cos \theta \ddot{\theta} - a \sin \theta \dot{\theta}^2) = F.$$

После группировки получаем первое уравнение движения системы:

$$(M + m)\ddot{x} + ma \cos \theta \ddot{\theta} = F + ma \sin \theta \dot{\theta}^2. \quad (6)$$

Для получения второго уравнения рассмотрим вращательное движение стержня. Запишем уравнение моментов относительно точки крепления (оси шарнира). Момент инерции стержня относительно конца равен  $I = \frac{1}{3}ml^2 = \frac{4}{3}ma^2$ . Уравнение вращательного движения:

$$ma \cos \theta \ddot{x} + I\ddot{\theta} = -mga \sin \theta. \quad (7)$$

### 3.1.3. Разрешение относительно старших производных

Объединим уравнения (6) и (7) в матричную форму относительно вектора угловых и линейных ускорений  $[\ddot{x}, \ddot{\theta}]^T$ :

$$\begin{pmatrix} M+m & ma \cos \theta \\ ma \cos \theta & I \end{pmatrix} \begin{pmatrix} \ddot{x} \\ \ddot{\theta} \end{pmatrix} = \begin{pmatrix} F + ma \sin \theta \dot{\theta}^2 \\ -mga \sin \theta \end{pmatrix}. \quad (8)$$

Определитель матрицы системы равен:

$$\Delta = (M + m)I - (ma \cos \theta)^2. \quad (9)$$

Решая систему методом Крамера или путем обратной матрицы, получаем явные выражения для ускорений, необходимые для численного моделирования:

$$\ddot{x} = \frac{I(F + ma \sin \theta \dot{\theta}^2) + m^2ga^2 \cos \theta \sin \theta}{\Delta}, \quad (10)$$

$$\ddot{\theta} = \frac{-(ma \cos \theta)(F + ma \sin \theta \dot{\theta}^2) - (M + m)mga \sin \theta}{\Delta}. \quad (11)$$

Полученная система уравнений (10)–(11) позволяет сформировать

**вектор состояния** системы  $s = [x, \dot{x}, \theta, \dot{\theta}]^T$ , который будет использоваться для реализации среды обучения с подкреплением.

### 3.1.4. Преобразование динамической системы в форму первого порядка

Исходная математическая модель, основанная на законах Ньютона, представляет собой систему обыкновенных дифференциальных уравнений (ОДУ) **второго порядка**, поскольку она выражается через вторые производные координат (ускорения  $\ddot{x}$  и  $\ddot{\theta}$ ).

Большинство численных методов интегрирования (например, метод Рунге-Кутты) разработаны для работы с системами ОДУ **первого порядка**. Для приведения системы к требуемой форме мы используем **вектор состояния**  $s$ , который включает все обобщенные координаты и их первые производные (скорости):

$$s = \begin{pmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{pmatrix}. \quad (12)$$

Производная этого вектора  $\dot{s}$  представляет собой систему первого порядка:

$$\dot{s} = \frac{d}{dt}s = \begin{pmatrix} \dot{x} \\ \ddot{x} \\ \dot{\theta} \\ \ddot{\theta} \end{pmatrix}. \quad (13)$$

Таким образом, динамическая система принимает форму  $\dot{s} = g(s, F)$ , где функция  $g(s, F)$  определяется:

$$g(s, F) = \begin{pmatrix} \dot{x} \\ \ddot{x}(s, F) \\ \dot{\theta} \\ \ddot{\theta}(s, F) \end{pmatrix}, \quad (14)$$

где  $\ddot{x}(s, F)$  и  $\ddot{\theta}(s, F)$  вычисляются по явным формулам (10) и (11).

### 3.1.5. Достаточность вектора состояния для динамики и RL

Для обучения с подкреплением агент оперирует **вектором наблюдения** (`observation_space`), который в данном случае совпадает с вектором состояния системы:

$$\mathbf{s}_{RL} = [x, \dot{x}, \theta, \dot{\theta}]^T. \quad (15)$$

**1. Достаточность для динамики (Марковское свойство):** Система описывается обыкновенными дифференциальными уравнениями (ОДУ) второго порядка. Для однозначного прогнозирования траектории системы в будущем необходимо знать текущее положение и скорость по каждой степени свободы. Вектор  $\mathbf{s}_{RL}$  включает все обобщенные координаты  $(x, \theta)$  и их первые производные  $(\dot{x}, \dot{\theta})$ , что обеспечивает выполнение **Марковского свойства**: будущее состояние системы полностью определяется текущим состоянием  $S_t$  и управляющим воздействием  $F_t$ .

**2. Пригодность для RL:** Вектор  $\mathbf{s}_{RL}$  является **вектором наблюдения** (`Observation`), который предоставляет агенту всю информацию, необходимую для принятия оптимальных решений:

- $x, \dot{x}$ : Необходимы для удержания тележки в пределах рабочей зоны.
- $\theta, \dot{\theta}$ : Являются критическими параметрами. Агент использует  $\theta$  для оценки отклонения от вертикали и  $\dot{\theta}$  для прогнозирования будущей траектории стержня (гашение колебаний и стабилизация).

### 3.1.6. Численное интегрирование методом Рунге-Кутты 4-го порядка

Для симуляции движения системы (переход  $s_t \rightarrow s_{t+\Delta t}$ ) используется **Метод Рунге-Кутты 4-го порядка (RK4)**. Этот метод обеспечивает

высокий порядок точности ( $O(\Delta t^4)$  на шаге) и хорошую устойчивость при моделировании нелинейных динамических систем.

RK4 является одношаговым методом, использующим четыре оценки наклона (производной  $g(s, F)$ ) для вычисления приращения вектора состояния на временном интервале  $\Delta t$ .

Пусть  $s_t$  — состояние в момент времени  $t$ ,  $F$  — управляющая сила, и  $g(s, F)$  — функция правой части системы  $\dot{s}$ . Шаг интегрирования  $\Delta t$ :

1.  $K_1$  (**Наклон в начале интервала**):

$$\mathbf{K}_1 = \Delta t \cdot g(s_t, F)$$

2.  $K_2$  (**Наклон в середине, с использованием  $K_1$** ):

$$\mathbf{K}_2 = \Delta t \cdot g(s_t + \mathbf{K}_1/2, F)$$

3.  $K_3$  (**Уточненный наклон в середине, с использованием  $K_2$** ):

$$\mathbf{K}_3 = \Delta t \cdot g(s_t + \mathbf{K}_2/2, F)$$

4.  $K_4$  (**Наклон в конце интервала, с использованием  $K_3$** ):

$$\mathbf{K}_4 = \Delta t \cdot g(s_t + \mathbf{K}_3, F)$$

5. **Итоговое состояние  $s_{t+\Delta t}$ :**

$$s_{t+\Delta t} = s_t + \frac{1}{6}(\mathbf{K}_1 + 2\mathbf{K}_2 + 2\mathbf{K}_3 + \mathbf{K}_4). \quad (16)$$

Этот численный метод реализуется в функции `step` симуляционной среды, обеспечивая эволюцию состояния системы под действием управляющей силы  $F$  и внутренних физических законов.

### 3.1.7. Функция вознаграждения

Функция вознаграждения  $r_t$  (Reward Function) определяет цель для RL-агента, поощряя желательное поведение (стабилизация) и наказывая нежелательное (падение или выход за пределы). На каждом времен-

ном шаге  $t$  вознаграждение рассчитывается по следующей формуле:

$$r = 1 - \left( C_1\theta^2 + C_2x^2 + C_3\dot{\theta}^2 + C_4\dot{x}^2 \right), \quad (17)$$

В данной реализации используются следующие весовые коэффициенты (константы):  $C_1 = 0.1$ ,  $C_2 = 0.1$ ,  $C_3 = 0.01$ ,  $C_4 = 0.01$ .

Базовое вознаграждение  $r$  стремится к максимальному значению  $r = 1$  при идеальном состоянии ( $\theta = 0, x = 0, \dot{\theta} = 0, \dot{x} = 0$ ). Каждый компонент в скобках представляет собой "штраф", который вычитается из единицы:

1. Штраф за положение маятника и тележки ( $0.1\theta^2 + 0.1x^2$ ): \* Наибольший вес (0.1) имеют штрафы, зависящие от **положения** — угла  $\theta$  и координаты  $x$ . Это заставляет агента фокусироваться на удержании маятника как можно ближе к вертикали ( $\theta = 0$ ) и тележки как можно ближе к центру ( $x = 0$ ). 2. Штраф за скорости ( $0.01\dot{\theta}^2 + 0.01\dot{x}^2$ ): \* Меньший вес (0.01) назначен штрафам, зависящим от **скоростей** ( $\dot{\theta}, \dot{x}$ ). Это поощряет агента к стабилизации системы (гашению скоростей), но не делает это приоритетом перед удержанием положений.

Если система выходит за установленные физические границы, налагается крупный единовременный штраф, и эпизод завершается:

$$\text{Если } (|x| > x_{\text{fail}} \text{ или } |\theta| > \theta_{\text{fail}}) : \quad r_{\text{fail}} = r - 10. \quad (18)$$

\*  $x_{\text{fail}}$  и  $\theta_{\text{fail}}$  — это предельные значения для координаты тележки и угла маятника, при превышении которых симуляция считается неудачной. \* Штраф в  $-10$  является значительным отрицательным вознаграждением, которое эффективно обучает агента избегать состояний, приводящих к сбою.

### 3.2. Вторая задача

Для моделирования динамики системы «Перевернутый маятник на тележке» и создания среды для обучения с подкреплением (RL) разработан класс `Model` на языке Python. Этот класс инкапсулирует математическую модель, численный решатель и логику взаимодействия,

соответствующую стандарту Gym.

### 3.2.1. Структура и параметры модели

Класс инициализируется с основными физическими и симуляционными параметрами:

- Массы:  $M = 1.2$  кг (тележка),  $m = 0.5$  кг (маятник).
- Длина маятника:  $l = 1.0$  м.
- Шаг интегрирования:  $dt = 0.01$  с.
- Максимальная управляющая сила:  $F_{\max} = 10$  Н.
- Границы сбоя:  $\theta_{\text{fail}} = \pi/2$  рад.,  $x_{\text{fail}} = 10.0$  м.

### 3.2.2. Функция динамики и интегрирование

Ключевыми компонентами являются методы, реализующие динамические уравнения и численный переход.

**Функция динамики  $g(s, F)$ :** Метод `g(self, state, action)` вычисляет производную вектора состояния  $\dot{S}$ , используя явные формулы для ускорений  $\ddot{x}$  и  $\ddot{\theta}$  (уравнения (10) и (11)). Для повышения устойчивости расчетов в знаменателе ( $\Delta$ ) реализована **численная стабилизация** с ограничением  $\max(\Delta, 10^{-4})$ , что предотвращает ошибку деления на ноль.

**Численное интегрирование:** Метод `apply_action(self, state, action, dt)` использует **Метод Рунге-Кутты 4-го порядка (RK4)** для перехода от состояния  $s_t$  к  $s_{t+dt}$ .

### 3.2.3. Интерфейс симуляции (Метод `step`)

Метод `step(self, state, action)` является основным интерфейсом взаимодействия RL-агента со средой. Он принимает текущее состояние и выбранную агентом силу  $F$  (`action`), после чего возвращает новое состояние, вознаграждение и флаг завершения эпизода (`done`).

Ниже представлен код, реализующий динамическую модель:

## Листинг 1: Реализация динамической модели «стержень на тележке»

---

```
1 import numpy as np
2
3 class Model():
4     def __init__(self, episode_length=1000, Fmax=10, dt=0.01):
5         self.l = 1.0
6         self.m = 0.5
7         self.M = 1.2
8         self.tetta_fail = np.pi/2
9         self.x_fail = 10.0
10        self.Fmax = Fmax
11        self.dt = dt
12        self.episode_length = episode_length
13
14    def step(self, state, action):
15        action = np.clip(action, -self.Fmax, self.Fmax)
16        new_state = self.apply_action(state, action, self.dt)
17        reward = self.reward(new_state)
18        done = abs(new_state[0]) > self.x_fail or abs(new_state[2]) > self.tetta_fail
19        return new_state, reward, done
20
21    def g(self, state, action):
22        x, x_der, tetta, tetta_der = state
23        g = 9.8
24        I = self.m * self.l**2 / 3
25        a = self.l / 2
26        # Вычисление определятеля Delta (знаменатель)
27        delta = I * (self.M + self.m) - (self.m * a * np.cos(tetta))**2
28        delta = max(delta, 1e-4) # Численная стабилизация
29        x_2der = (I * (action + self.m * a * tetta_der**2 * np.sin(tetta)) +
30                  self.m**2 * g * a**2 * np.cos(tetta) * np.sin(tetta)) / delta
31        tetta_2der = (-self.m * a * np.cos(tetta) * (action + self.m * a * tetta_der**2 *
32                           np.sin(tetta)) -
33                           (self.M + self.m) * self.m * g * a * np.sin(tetta)) / delta
34        return np.array([x_der, x_2der, tetta_der, tetta_2der])
35
36    def apply_action(self, state, action, dt):
37        # Метод Рунге-Куммы 4-го порядка (RK4)
38        k1 = self.g(state, action)
39        k2 = self.g(state + dt * 0.5 * k1, action)
40        k3 = self.g(state + dt * 0.5 * k2, action)
41        k4 = self.g(state + dt * k3, action)
42        new_state = state + dt * (1/6 * (k1 + 2*k2 + 2*k3 + k4))
```

```

42     return new_state
43
44     def reward(self, state):
45         x, x_der, tetta, tetta_der = state
46         # Базовое вознаграждение
47         r = 1 - (0.1 * tetta ** 2 + 0.1 * x ** 2 + 0.01 * tetta_der ** 2 + 0.01 * x_der
48             ** 2)
49         # Дополнительный штраф за сбой
50         if abs(x) > self.x_fail or abs(tetta) > self.tetta_fail:
51             r -= 10
52     return r

```

---

### 3.3. Третья задача

В данном разделе представлена программная реализация трех выбранных алгоритмов обучения, предназначенных для поиска оптимальной политики стабилизации стержня на тележке (CartPole).

#### 3.3.1. Реализация Basic Random Search (BRS)

**Basic Random Search (BRS)** — это простейший представитель безградиентных методов оптимизации политики. В контексте задачи CartPole, целью BRS является поиск оптимального вектора весов  $\mathbf{w}$  для **линейной политики**, которая связывает вектор состояния  $\mathbf{s} = [x, \dot{x}, \theta, \dot{\theta}]^T$  с управляющей силой  $F$ :

$$F = \mathbf{w}^T \mathbf{s} = w_1 x + w_2 \dot{x} + w_3 \theta + w_4 \dot{\theta}$$

**Принцип работы BRS** основан на глобальном случайному поиске:

- **Генерация:** На каждой итерации (эпизоде) генерируется совершенно новый вектор весов  $\mathbf{w}$  из равномерного распределения в заданном диапазоне.
- **Оценка:** Производится прогон модели (метод `rollout`) с использованием новой политики  $\mathbf{w}$ , и вычисляется суммарная награда  $r$ .

- **Обновление:** Если полученная награда  $r$  превышает текущую наилучшую награду ( $r_{best}$ ), то  $\mathbf{w}$  сохраняется как новая лучшая политика  $\mathbf{w}_{best}$ .

Этот метод обеспечивает высокую вероятность выхода из локальных оптимумов, но его сходимость крайне низка, что делает его идеальным **базовым алгоритмом** для сравнения.

**Детали программной реализации** Класс `BRSagent` инкапсулирует логику поиска.

- Метод `learn` реализует основной цикл, где ключевая строка `self.w = np.random.uniform(-1, 1, size=4)` обеспечивает **полностью случайную генерацию** нового кандидата на каждой итерации.
- Метод `rollout` симулирует эпизод, используя линейную политику `action = np.clip(np.dot(w, state), ...)` и гарантирует, что управляющая сила  $F$  не превышает физического ограничения  $F_{max}$ .
- Метод `reset_state` инициализирует начальное состояние маятника  $\mathbf{s}_0$  в случайной области.

## Исходный код Basic Random Search (BRS)

## Листинг 2: Реализация Basic Random Search

---

```
1 import numpy as np
2 from src.Model import Model
3
4 class BRSagent():
5     def __init__(self, model=Model(), n_episodes=100):
6         # Инициализация случайными весами (линейная политика)
7         self.w = np.random.uniform(-1, 1, size=4)
8         self.n_episodes = n_episodes
9         self.model = model
10        self.best_w = self.w.copy()
11        self.best_reward = -np.inf
12        self.states = []
13        self.actions = []
14        self.rewards = []
15
16    def learn(self):
17        for episode_ind in range(self.n_episodes):
18            # Полностью новый случайный вектор весов
19            self.w = np.random.uniform(-1, 1, size=4)
20            r = self.rollout(self.w)
21            # Сохранение лучшей политики
22            if r > self.best_reward:
23                self.best_reward = r
24                self.best_w = self.w.copy()
25
26    def rollout(self, w):
27        state = self.reset_state()
28        total_reward = 0
29        done = False
30        steps = 0
31        while not done and steps < self.model.episode_length:
32            # Линейная политика с ограничением F_max
33            action = np.clip(np.dot(w, state), -self.model.Fmax, self.model.Fmax)
34            new_state, reward, done = self.model.step(state, action)
35            self.states.append(state)
36            self.actions.append(action)
37            self.rewards.append(reward)
38            state = new_state
39            total_reward += reward
40            steps += 1
41        return total_reward
42
43    def reset_state(self):
```

```

44     # Инициализация состояния в случайном диапазоне
45     self.states = []
46     self.actions = []
47     self.rewards = []
48     state = np.array([
49         np.random.uniform(-1.0, 1.0),          # x
50         np.random.uniform(-0.1, 0.1),          # x_dot
51         np.random.uniform(-np.pi/6, np.pi/6), # tetta
52         np.random.uniform(-0.1, 0.1)          # tetta_dot
53     ], dtype=float)
54     return state

```

---

### 3.3.2. Реализация Hill Climbing (HC)

**Hill Climbing (HC)** (Восхождение на вершину) является итеративным алгоритмом локального поиска, который представляет собой усовершенствование BRS за счет целенаправленной эксплуатации окрестности уже найденного лучшего решения. В отличие от BRS, HC не генерирует полностью новую политику на каждой итерации, а вводит небольшое случайное **возмущение** ( $\delta$ ) к текущему лучшему вектору весов  $\mathbf{w}_{\text{best}}$ , формируя политику-кандидат  $\mathbf{w}'$ :

$$\mathbf{w}' = \mathbf{w}_{\text{best}} + \delta, \quad \text{где } \delta \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

Здесь  $\delta$  — вектор случайного шума, распределенного по Гауссу со стандартным отклонением  $\sigma$  (параметр, контролирующий радиус локального поиска).

**Принцип работы HC:**

- **Генерация кандидата:** Политика-кандидат  $\mathbf{w}'$  генерируется путем добавления гауссовского шума к текущему лучшему вектору весов  $\mathbf{w}_{\text{best}}$ .
- **Оценка:** Производится прогон модели (`rollout`) с использованием  $\mathbf{w}'$ , и вычисляется суммарная награда  $r$ .
- **Обновление:** Если полученная награда  $r$  превышает текущую наилучшую награду ( $r_{\text{best}}$ ), то  $\mathbf{w}'$  становится новой  $\mathbf{w}_{\text{best}}$ . Если улучшения нет,  $\mathbf{w}_{\text{best}}$  остается неизменным.

Данная реализация включает механизм **затухания**  $\sigma$  (`sigma_decay`), который позволяет уменьшать радиус поиска с течением времени. Это способствует переходу от широкого исследования пространства в начале обучения к точной настройке (эксплуатации) найденного оптимума на более поздних стадиях. НС, как правило, сходится быстрее, чем BRS, но более подвержен риску застревания в локальных оптимумах.

**Детали программной реализации** Класс `HCagent` включает два новых ключевых параметра: `sigma` и `sigma_decay`.

- В методе `learn` генерируется возмущение:  
`delta = np.random.normal(0, self.sigma, size=4)`, а затем вычисляется кандидат `new_w = self.w + delta`.
- Обновление весов происходит только при улучшении, что соответствует классическому алгоритму НС.
- В конце каждой итерации  $\sigma$  уменьшается:  
`self.sigma = max(self.sigma * self.sigma_decay, 0.01)`.

## Исходный код Hill Climbing (НС)

### Листинг 3: Реализация Hill Climbing (HC) с затуханием шума

---

```
1 import numpy as np
2 from src.Model import Model
3
4 class HCagent():
5     def __init__(self, model=Model(), sigma=0.5, sigma_decay=0.995, n_episodes=100):
6         self.w = np.random.uniform(-1, 1, size=4)
7         self.best_w = self.w.copy()
8         self.sigma = sigma                      # Параметр шума (радиус поиска)
9         self.sigma_decay = sigma_decay          # Коэффициент затухания шума
10        self.n_episodes = n_episodes
11        self.model = model
12        self.best_reward = -np.inf
13        self.states = []
14        self.actions = []
15        self.rewards = []
16
17    def learn(self):
18        for episode_ind in range(self.n_episodes):
19            # Генерация гауссского шума (дельта)
20            delta = np.random.normal(0, self.sigma, size=4)
21            # Новый кандидат w'
22            new_w = self.w + delta
23            reward = self.rollout(new_w)
24            # Обновление w только при улучшении (восхождение)
25            if reward > self.best_reward:
26                self.w = new_w
27                self.best_reward = reward
28                self.best_w = new_w.copy()
29            # Затухание радиуса поиска
30            self.sigma = max(self.sigma * self.sigma_decay, 0.01)
31
32    def rollout(self, w):
33        state = self.reset_state()
34        total_reward = 0
35        done = False
36        steps = 0
37        while not done and steps < self.model.episode_length:
38            # Линейная политика с ограничением F_max
39            action = np.clip(np.dot(w, state), -self.model.Fmax, self.model.Fmax)
40            new_state, reward, done = self.model.step(state, action)
41            self.states.append(state)
42            self.actions.append(action)
```

```

43         self.rewards.append(reward)
44         state = new_state
45         total_reward += reward
46         steps += 1
47
48     return total_reward
49
50
51     def reset_state(self):
52         # Инициализация состояния в случайном диапазоне
53         self.states = []
54         self.actions = []
55         self.rewards = []
56         state = np.array([
57             np.random.uniform(-1.0, 1.0),           # x
58             np.random.uniform(-0.1, 0.1),          # x_dot
59             np.random.uniform(-np.pi/6, np.pi/6),   # tetta
60             np.random.uniform(-0.1, 0.1)           # tetta_dot
61         ], dtype=float)
62
63     return state

```

---

### 3.3.3. Реализация архитектуры Actor-Critic (AC)

В отличие от BRS и НС, метод **Actor-Critic** относится к классу алгоритмов обучения с подкреплением, использующих градиентную оптимизацию. Это гибридная архитектура, сочетающая в себе преимущества методов на основе ценности (Value-based) и методов на основе политики (Policy-based).

Архитектура состоит из двух линейных аппроксиматоров:

1. **Actor (Актер):** Определяет стохастическую политику  $\pi_\theta(a|s)$ , выбирая действие  $a$  в состоянии  $s$ . Его цель — максимизировать ожидаемую награду.
2. **Critic (Критик):** Оценивает функцию ценности состояния  $V_w(s)$ , предсказывая ожидаемую сумму будущих наград. Его цель — минимизировать ошибку предсказания (TD-ошибку).

**Математическая модель:** В данной реализации используется параметризация политики Гауссовским распределением:  $a \sim \mathcal{N}(\mu(s), \sigma(s))$ . Параметры распределения ( $\mu$  и  $\sigma$ ) и функция ценности  $V(s)$  аппроксируются линейными моделями от вектора состояния  $s$ :

$$\mu(s) = \theta_\mu^T \mathbf{s}, \quad \sigma(s) = \exp(\theta_\sigma^T \mathbf{s}), \quad V(s) = \mathbf{w}^T \mathbf{s}$$

Обновление весов происходит на каждом шаге (online-обучение) на основе ошибки временной разности (TD-error), которая показывает, насколько действие оказалось лучше или хуже ожидаемого:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

Градиенты для обновления параметров Актера выводятся через теорему о градиенте политики (Policy Gradient Theorem) и логарифмическую производную функции плотности вероятности нормального распределения.

**Детали программной реализации** Класс `ACagent` содержит параметры для обучения ( $\alpha_{actor}, \alpha_{critic}, \gamma$ ) и весовые коэффициенты: `w` (для Критика), `tetta_mu` и `tetta_sigma` (для Актера).

Ключевые методы реализации:

- **Расчет TD-ошибки:** В методе `learn` вычисляется `TD_error`, которая выступает «учителем» для Актера. Если ошибка положительна, вероятность выбранного действия увеличивается, если отрицательна — уменьшается.
- **Обновление Критика:** Веса `w` обновляются методом стохастического градиентного спуска, чтобы  $V(s)$  приближалась к реальной полученной награде.
- **Обновление Актера:** Используются функции `acceptance_mu` и `acceptance_sigma`, которые вычисляют градиенты  $\nabla_\theta \ln \pi(a|s)$ . Для  $\mu$  градиент пропорционален  $\frac{a-\mu}{\sigma^2}$ , а для  $\sigma$  (учитывая экспоненциальную параметризацию) —  $(\frac{(a-\mu)^2}{\sigma^2} - 1)$ .
- **Клиппинг (Clipping):** Для обеспечения стабильности обучения и предотвращения «взрыва градиентов», параметры  $\sigma$  и сами весовые коэффициенты принудительно ограничиваются в заданных диапазонах (`np.clip`).

## **Исходный код Actor-Critic**

## Листинг 4: Реализация алгоритма Actor-Critic

---

```
1 import numpy as np
2 from src.Model import Model
3
4 class PGagent():
5     def __init__(self, lr_actor=1e-3, lr_critic=1e-2, gamma=0.99, model=Model(),
6                  n_episodes=100):
7         self.lr_actor = lr_actor      # Скорость обучения Актера
8         self.lr_critic = lr_critic    # Скорость обучения Критика
9         self.gamma = gamma           # Коэффициент дисконтирования
10        self.model = model
11        self.n_episodes = n_episodes
12        # Инициализация весов
13        self.w = np.random.uniform(-1, 1, size=4)          # Веса Критика (Value func)
14        self.tetta_mu = np.zeros(4)                         # Веса Актера (Мат. ожидание)
15        self.tetta_sigma = np.ones(4) * 0.1                 # Веса Актера (Дисперсия)
16        # Параметры ограничения (clipping) для стабильности
17        self.sigma_clip_min = 0.05
18        self.sigma_clip_max = 2.0
19        self.tetta_sigma_clip = 5.0
20        self.tetta_mu_clip = 10
21        self.states = []
22        self.actions = []
23        self.rewards = []
24
25    def learn(self):
26        for episode_ind in range(self.n_episodes):
27            self.rollout()
28            # Пошаговое обновление весов (Online learning)
29            for i in range(len(self.actions)):
30                action, state, reward = self.actions[i], self.states[i], self.rewards[i]
31                # Оценка ценности следующего состояния  $V(s')$ 
32                next_value = self.value_func(self.states[i+1]) if i+1 < len(self.states)
33                else 0
34                # Вычисление ошибки временной разности:  $\delta = r + \gamma V(s') - V(s)$ 
35                TD_error = reward + self.gamma * next_value - self.value_func(state)
36                # Обновление весов Актера (Policy Update)
37                # acceptance_... реализуют градиент логарифма правдоподобия
38                self.tetta_mu += self.lr_actor * TD_error * self.acceptance_mu(action,
→ state)
                self.tetta_sigma += self.lr_actor * TD_error *
→ self.acceptance_sigma(action, state)
                # Ограничение весов для предотвращения расходности
```

```

39         self.tetta_sigma = np.clip(self.tetta_sigma, -self.tetta_sigma_clip,
40             ↵ self.tetta_sigma_clip)
41         self.tetta_mu = np.clip(self.tetta_mu, -self.tetta_mu_clip,
42             ↵ self.tetta_mu_clip)
43         # Обновление весов Критика (Value Update)
44         # Градиент  $V(s)$  по  $w$  равен самому вектору состояния  $state$  (так как  $V$ 
45             ↵ линейна)
46         self.w += self.lr_critic * TD_error * state
47
48
49     def rollout(self):
50         state = self.reset_state()
51         total_reward = 0
52         done = False
53         steps = 0
54         while not done and steps < self.model.episode_length:
55             # Выбор действия стochастической политикой
56             mu_val = self.mu(state)
57             sigma_val = self.sigma(state)
58             action = self.policy(mu_val, sigma_val)
59             new_state, reward, done = self.model.step(state, action)
60             self.states.append(state)
61             self.actions.append(action)
62             self.rewards.append(reward)
63             state = new_state
64             total_reward += reward
65             steps += 1
66
67     return total_reward
68
69
70     def value_func(self, state):
71         # Линейная аппроксимация  $V(s) = w^T * s$ 
72         return np.dot(state, self.w)
73
74
75     def policy(self, mu, sigma):
76         # Сэмплирование действия из нормального распределения
77         a = np.random.normal(mu, sigma)
78         return np.clip(a, -self.model.Fmax, self.model.Fmax)
79
80
81     def sigma(self, s):
82         # Вычисление стандартного отклонения:  $\sigma = \exp(\theta_\sigma \cdot T * s)$ 
83         z = np.dot(s, self.tetta_sigma)
84         z = np.clip(z, -10, 10) # Защита от переполнения экспоненты
85         sigma = np.clip(np.exp(z), self.sigma_clip_min, self.sigma_clip_max)
86         return sigma
87
88
89     def mu(self, s):
90         # Вычисление мат. ожидания:  $\mu = \theta_\mu \cdot T * s$ 
91         mu = np.dot(s, self.tetta_mu)

```

```

83     return mu
84
85     def acceptance_mu(self, action, state):
86         # Градиент log-probability по параметрам mu
87         # d(ln pi)/d(mu) * d(mu)/d(theta_mu)
88         acceptance = state * (action - self.mu(state)) / (self.sigma(state) ** 2 + 1e-8)
89         return acceptance
90
91     def acceptance_sigma(self, action, state):
92         # Градиент log-probability по параметрам sigma
93         # Учитываем, что sigma параметризована через экспоненту
94         acceptance = state * ((action - self.mu(state)) ** 2 / (self.sigma(state) ** 2 +
95             ↪ 1e-8) - 1)
96         return acceptance
97
98     def reset_state(self):
99         self.states = []
100        self.actions = []
101        self.rewards = []
102        state = np.array([
103            np.random.uniform(-1.0, 1.0),           # x
104            np.random.uniform(-0.1, 0.1),           # x_dot
105            np.random.uniform(-np.pi/6, np.pi/6),   # tetta
106            np.random.uniform(-0.1, 0.1)           # tetta_dot
107        ], dtype=float)
108        return state

```

---

## 4. Эксперимент

Как мы проверяем, что всё удачно получилось. Если работа рассчитана на несколько семестров и в текущем до эксперимента дело не дошло, опишите максимально подробно, как он будет проводиться и на чём (то, что называется *дизайн эксперимента* — от того, что и как Вы будете проверять, очень сильно зависит, что Вы будете делать, так что это важно и делается *не* после реализации).

### 4.1. Условия эксперимента

Железо (если актуально); версии ОС, компиляторов и параметры командной строки; почему мы выбрали именно эти тесты; входные данные, на которых проверяем наш подход, и почему мы выбрали именно их.

### 4.2. Исследовательские вопросы

По-английски называется *research questions*, в тексте можно ссылаться на них как RQ1, RQ2, и т. д. Необходимо сформулировать, чего мы хотели бы добиться работой (2 пункта будет хорошо):

**RQ1** : правда ли предложенный в работе алгоритм лучше вот таких-то остальных?

**RQ2** : насколько существенно каждая составляющая влияет на улучшения? (Если в подходе можно включать/выключать какие-то составляющие.)

**RQ3** : насколько точны полученные приближения, если работа строит приближения каких-то штук?

и т.п.

Иногда в работах это называют гипотезами, которые потом проверяют. Далее в тексте можно ссылаться на исследовательские вопросы как RQ, это общепринятое сокращение.

## 4.3. Метрики

Как мы сравниваем, что результаты двух подходов лучше или хуже:

- Производительность.
- Строчки кода.
- Как часто алгоритм «угадывает» правильную классификацию входа.

Иногда метрики вырожденные (да/нет), это не очень хорошо, но если в области исследований так принято, то ладно. Если метрики хитрые (даже IoU или  $F_1$ -меру можно считать хитрыми), разберите их в обзоре, пояснив, почему выбраны именно такие метрики.

## 4.4. Результаты

Результаты понятно что такое. Тут всякие таблицы и графики, как в таблице 3. Обратите внимание, как цифры выровнены по правому краю, названия по центру, а разделители  $\times$  и  $\pm$  друг под другом.

Перед написанием данного раздела имеет смысл проконсультироваться с литературой по проведению экспериментов [5].

Скорее всего Ваши измерения будут удовлетворять нормальному распределению, в идеале это надо проверять с помощью критерия Колмогорова и т.п. Если критерий этого не подтверждает, то у Вас что-то сильно не так с измерениями, надо проверять кэши процессора, отключать Интернет во время измерений, подкручивать среду исполнения (англ. runtime), чтобы сборка мусора не вмешивалась и т.п. Если критерий удовлетворён, то необходимо либо указать мат. ожидание и доверительный/предсказывающий интервал, либо мат. ожидание и среднеквадратичное отклонение, либо, если совсем лень заморачиваться, написать, что все измерения проводились с погрешностью, например, в 5%. Не приводите слишком много значащих цифр (например, время работы в 239.1 секунды при среднеквадратичном отклонении в

**Таблица 3:** Производительность какого-то алгоритма при различных разрешениях картинок (меньше — лучше), в мс., CI=0.95. За пример таблички кидаем чепчики в честь Я. Кириленко

Resolution	TENG	LAPM	VOLL4
$1920 \times 1080$	$406.23 \pm 0.94$	$134.06 \pm 0.35$	$207.45 \pm 0.42$
$1024 \times 768$	$145.00 \pm 0.47$	$39.68 \pm 0.10$	$52.79 \pm 0.10$
$464 \times 848$	$70.57 \pm 0.20$	$19.86 \pm 0.01$	$32.75 \pm 0.04$
$640 \times 480$	$51.10 \pm 0.20$	$14.70 \pm 0.10$	$24.00 \pm 0.04$
$160 \times 120$	$2.40 \pm 0.02$	$0.67 \pm 0.01$	$0.92 \pm 0.01$

50 секунд выглядит глупо, даже если ваш любимый бенчмарк так посчитал).

Замечание: если у вас получится улучшение производительности в пределах погрешности, то это обязательно вызовет вопросы. Если погрешность получилась значительной (больше 10-15% от среднего), это тоже вызовет вопросы, на которые надо ответить, либо разобравшись, что не так (первый подозреваемый — мультимодальное распределение), либо более глубоко статистически проанализировав результаты (например, привести гистограммы).

В этом разделе надо также явно ответить на Research Questions или как-то их прокомментировать.

#### 4.4.1. RQ1

Пояснения

#### 4.4.2. RQ2

Пояснения

## 4.5. Обсуждение результатов

Чуть более неформальное обсуждение, то, что сделано. Например, почему метод работает лучше остальных? Или, что делать со случаями, когда метод классифицирует вход некорректно.

## 4.6. Угрозы нарушения корректности (опциональный)

Если основная заслуга метода, это то, что он дает лучшие цифры, то стоит сказать, где мы могли облажаться, когда

1. проводили численные замеры;
2. выбирали тестовый набор (см. *confirmation bias*).

Например, если мы делали замер юзабилити нашего продукта по методике System Usability Scale, но в эксперименте участвовали Ваши друзья, которые хотят Вас порадовать, честно напишите об этом. Честно напишите, какие меры были приняты для минимизации рисков валидности результатов эксперимента и, если это содержательно, почему их не удалось полностью исключить.

## 4.7. Воспроизводимость эксперимента

Это настолько важно, что заслуживает тут своего подраздела (в самой работе не надо, это должно естественно вытекать из разделов выше) — эксперимент должно быть можно повторить и получить примерно такие же результаты, как у Вас. Поэтому выложите свой код, которым мерили. Выложите данные, на которых мерили. Или напишите, где эти данные взять. Напишите, как конкретно запустить, в каком окружении, что надо дополнительно поставить и т.п. Чтобы любой второкурсник мог выполнить все пункты и получить тот же график/таблицу, что у Вас. Не обязательно это делать прямо в тексте, можете в README своего репозитория, но где-то надо.

Код модуля в составе дисциплины, практики и т.п.	Трудоёмкость	Контактная работа обучающихся с преподавателем									Самостоятельная работа с использованием методических средств	Объем активных и интерактивных форм работы
		лекции	семинары	консультации	практические занятия	лабораторные работы	контрольные работы	коллоквиумы	текущий контроль	промежуточная аттестация		
Семестр 3	2	30								18	20	10
		2–42							2–25	1–1	1–1	
Итого	2	30							2	18	20	10

Таблица 4: Если таблица очень большая, то можно её изобразить на отдельной портретной странице. Не забудьте подробное описание, чтобы содержимое таблицы можно было понять не читая весь текст.

## Список литературы

- [1] A2C is a special case of PPO / Shengyi Huang, Anssi Kanervisto, Antonin Raffin et al. // arXiv preprint arXiv:2205.09123. — 2022.
- [2] Brooks Jr. Frederick P. The Mythical Man-month (Anniversary Ed.). — Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 1995. — ISBN: [0-201-83595-9](#).
- [3] Sutton Richard S., Barto Andrew G. Reinforcement Learning: An Introduction. — MIT Press, 1998. — ISBN: [0262193981](#).
- [4] Williams Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning // Machine Learning. — 1992. — Vol. 8, no. 3-4. — P. 229–256.
- [5] Смирнов К. К. Шпаргалка о проведении экспериментов. — URL: <https://github.com/spbu-se/measurements> (дата обращения: 16 января 2024 г.).