Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа Аэрокосмических Технологий
Кафедра математического моделирования и прикладной математики

Направление подготовки / специальность: 03.03.01 Прикладные математика и физика
Направленность (профиль) подготовки: Геокосмические науки и технологии

УПРАВЛЕНИЕ КОСМИЧЕСКИМ АППАРАТОМ В ОБЛАСТИ ГРАВИТАЦИОННОЙ ЛИНЗЫ СОЛНЦА С ПОМОЩЬЮ МЕТОДОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

(бакалаврская работа)

Студент:
Гончарова Алина Эдуардовна

_____
(подпись студента)

Научный руководитель:
Перепухов Денис Глебович,

_____
(подпись научного руководителя)

Консультант *(при наличии)*:
_____

_____
(подпись консультанта)

Москва 2025

Аннотация

Получение прямых изображений удаленных экзопланет представляет собой важную научную задачу, однако она затруднена фундаментальными техническими и физическими ограничениями традиционных телескопов. В качестве альтернативы учёные предложили использовать Солнце как гравитационную линзу, фокусирующую свет от удаленных объектов (например, экзопланет) на расстоянии от 550 а.е. Для получения изображений с высоким разрешением необходима точная автономная навигация космического аппарата в области фокуса. В данной работе исследуется применение алгоритмов обучения с подкреплением для управления аппаратом в области фокуса гравитационной линзы Солнца. Рассматривается задача в двух постановках: упрощённой, когда положение фокуса фиксировано в пространстве, и в более реалистичной, учитывающей сложное движение фокуса. Представлены архитектура алгоритма обучения с подкреплением и полученные в результате обучения стратегии управления. В обеих постановках удалось получить законы управления, справляющиеся с поставленной задачей управления.

Оглавление

Введение

В последние годы всё больше внимания уделяется идее использования гравитационной линзы Солнца для наблюдения далёких объектов во Вселенной, например, экзопланет. Согласно общей теории относительности, массивные тела, такие как, например, Солнце, за счёт своей гравитации могут преломлять свет подобно обычным линзам – так называемый, эффект гравитационного линзирования. При этом лучи света от одного и того же источника будут сходиться в области, называемой гравитационным фокусом. Если разместить телескоп в области гравитационного фокуса Солнца (ГФС), который начинается на расстоянии примерно 550 астрономических единиц (а.е.) от Солнца, то, теоретически, можно получить изображение удалённого объекта (до 100 световых лет), в том числе экзопланеты, с разрешением до 10 км на пиксель [1]. Получить изображения экзопланет с таким разрешением, используя традиционные телескопы, как наземные, так и космические, крайне проблематично, из-за технических трудностей, а также из-за большой удаленности планет.

Однако миссия по отправке телескопа в ГФС имеет ряд серьёзных трудностей, как технических, так и баллистических. Если рассматривать только этап работы аппарата в ГФС, то, во-первых, область ГФС расположена на расстоянии более 550 а.е. от Солнца, что эквивалентно 75.9 световых часов, поэтому связь с Землёй сможет осуществляться только с большой задержкой. Во-вторых, для получения изображения экзопланеты аппарату будет необходимо осуществлять управление и навигацию вблизи фокальной линии с точностью до единиц метров. С учётом большого расстояния от Земли, это ведёт к тому, что навигация и управление должны будут осуществляться полностью автономно. К этому добавляются трудности, связанные со сложной динамикой движения оси ГФС, с неопределённостями в модели этого движения и с отсутствием навигационного обеспечения с Земли.

Чтобы решить задачу управления аппаратом в таких условиях, можно использовать разные подходы, основанные на принципе максимума Понтрягина, теории устойчивости Ляпунова и т.д. Один из современных и активно развивающихся подходов к решению задач в условиях неопределенности — это обучение с подкреплением (ОП). Для применения методов ОП к задаче управления динамической системой функцию управления (как правило, с обратной связью по состоянию) задают с помощью некоторой функции с большим числом настраиваемых параметров, значения которых подбирают в процессе оптимизации («обучения»). При этом задача управления рассматривается как взаимодействие так называемого агента со средой, от которой агент за свои действия получает вознаграждения и стремится максимизировать суммарную награду. В динамике космического полета, например, агентом может быть программное обеспечение на борту

аппарата, а средой сам аппарат, функцией вознаграждения – точность прилета в заданную область. Результатом обучения будет функция управления, которая способна направлять аппарат в заданную область пространства. Эта функция может быть загружена на борт космического аппарата и может управлять им во время реального полета на основе состояния аппарата или оценок состояния аппарата.

Методы ОП уже применялись в задачах механики космического полета. Например, в [2] был предложен подход удержания аппарат на гало-орбите вокруг точки либрации L1 с помощью методов обучения с подкреплением. В работе [3] автор исследует, как с помощью ОП можно получить эффективную стратегию управления, использующую многообразия, возникающие в круговой ограниченной задачи трёх тел, для поддержания движения космического аппарата вблизи неустойчивых симметричных периодических орбит. В другой работе [4] с помощью ОП и прямого метода Ляпунова разрабатывалось управление для перелётов к Луне. Было изучено применение этих методов для поддержания неустойчивых гало-орбит в рамках модели круговой ограниченной задачи трех тел. Таже возможно использование методов обучения с подкреплением для создания алгоритмов управления по спасению миссий в окололунном пространстве при возникновении нештатной ситуации. Так, в работе [5] рассматривается задача автоматического перепланирования траектории космического аппарата с малой тягой после сбоя, в частности, при неисправной работе двигателя. Работа демонстрирует, как методы ОП могут быть использованы не только для построения первичного управления, но и как инструмент для принятия решений в условиях неопределенности и ограниченного времени. Что же касается применения методов ОП к задаче управления космическим аппаратом в области ГФС, то этот вопрос изучался в работе [6], однако, только в упрощённой постановке, в которой не учитывалось движение оси ГФС, вызванное движениями Солнца и экзопланеты. Более детальное исследование динамики и управления в ГФС представлено в работе [1], где рассматривается постановка задачи в неинерциальной системе отсчета, предлагаются различные алгоритмы управления (включая ПД-регулятор, компенсационное и квази оптимальное по времени управления, а также их модификации) и проводится их численное тестирование с учетом навигационных и модельных ошибок.

В данной работе исследуется применение методов обучения с подкреплением для управления космическим аппаратом в области гравитационного фокуса Солнца. Целью является разработка автономного управления с обратной связью по состоянию для начального этапа миссии в ГФС, когда аппарат нужно приблизить к оси ГФС, но при этом высокоточное позиционирование ещё не требуется. Вопрос осуществления навигации и оценки состояния не рассматривается. Задача решалась в двух постановках. В первой

постановке использовалась упрощённая модель движения без учёта движений Солнца и экзопланеты (как в работе [6]). Во второй – более сложная модель, учитывающая движения Солнца и экзопланеты.

В главе 2 приведена постановка задачи в терминах классической механики. Введены необходимые системы координат, приведены уравнения движения для каждой из постановок, а также сформулирована задача оптимального управления. В главе 3 осуществляется переход к формализации задачи в рамках парадигмы обучения с подкреплением. Описаны основные принципы ОП и алгоритм Proximal Policy Optimization (PPO), задача оптимального управления сведена к задаче ОП. В главе 4 представлены результаты применения ОП для получения законов управления космическим аппаратом в ГФС. Приведены параметры обучения, представлены результаты обучения в обоих постановках задачи, исследована переносимость алгоритма управления, полученного в результате обучения в первой постановке, на вторую постановку. В заключении подведены итоги проведённого исследования, сформулированы основные выводы и обозначены направления для будущих исследований.

1. Постановка задачи

1.1. Принцип работы гравитационной линзы Солнца

Предположим, что существует экзопланета, находящаяся на расстоянии z_0 световых лет от Солнечной системы (рис. 1). Это расстояние зависит от времени, однако в дальнейшем для упрощения модели будем считать его постоянным. Пока что будем рассматривать экзопланету как точку. Линию, проходящую через центр экзопланеты и через центр Солнца, будем называть фокальной линией ГФС. Следует обратить внимание на то, что фокальная линия определяется с помощью двух объектов: Солнца и экзопланеты. Строго говоря, каждой экзопланете (или точке на поверхности одной экзопланеты) соответствует своя фокальная линия. Далее, однако, мы опускаем эту деталь и определяем фокальную линию только для центра экзопланеты.

Экзопланета освещается своей звездой-хозяином, отражённый экзопланетой свет движется от нее к Солнцу. Когда свет проходит вблизи Солнца, гравитация отклоняет световые лучи, фокусируя их на фокальной линии. Конкретная точка на фокальной линии, через которую проходит преломлённый световой луч, зависит от расстояния, на котором луч пролетает мимо Солнца: чем больше расстояние, тем дальше точка. Поскольку световые лучи проходят на различных расстояниях от Солнца, но не ближе солнечного радиуса, общий свет от экзопланеты фокусируется на геометрическом луче, лежащем на фокальной линии и начинающемся на расстоянии примерно 550 а.е. от Солнца. Если рассматривать экзопланету как протяженный источник, то нужно учитывать, что каждая точка поверхности экзопланеты соответствует уникальной фокальной линии. В этом случае свет от экзопланеты фокусируется не на луч, а в его окрестности – эту окрестность мы и называем областью гравитационного фокуса Солнца (область ГФС).

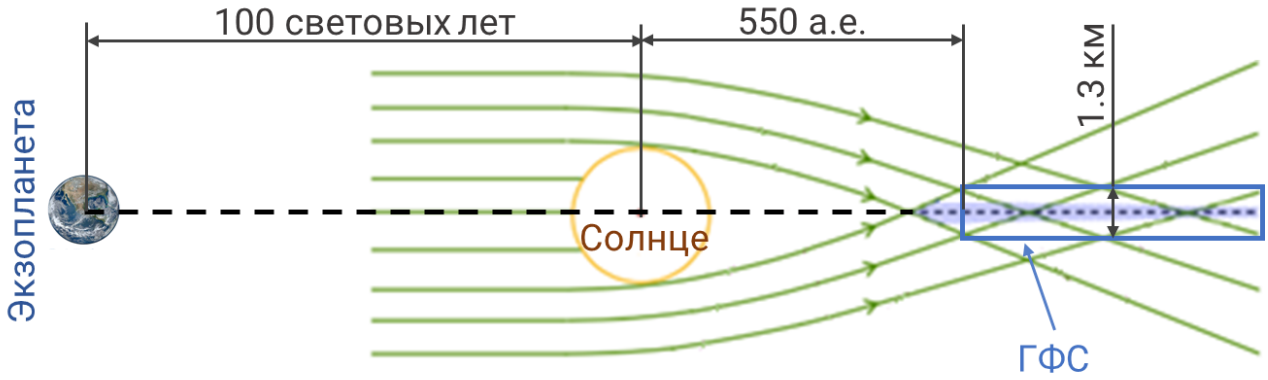

Рисунок 1. Схематическое представление солнечной гравитационной линзы из [1].

Для описания того, как формируется изображение экзопланеты, вводятся две плоскости. Плоскость источника перпендикулярна фокальной линии и проходит через

центр экзопланеты (рис. 2). Пренебрегая кривизной поверхности экзопланеты, предполагается, что весь свет от экзопланеты исходит из плоскости источника. Плоскость изображения — это любая плоскость, ортогональная фокальной линии и пересекающая ГФС. Эти плоскости позволяют удобно описать процесс формирования изображения: если z – это расстояние от Солнца до плоскости изображения, а $\mathbf{x}_s$ указывает положение точки в плоскости источника, то эта точка проецируется в соответствующую точку в плоскости изображения $\mathbf{x}_{im} = -\frac{z}{z_0}\mathbf{x}_s$. Таким образом строится биекция между точками в плоскости источника и точками в плоскости изображения.

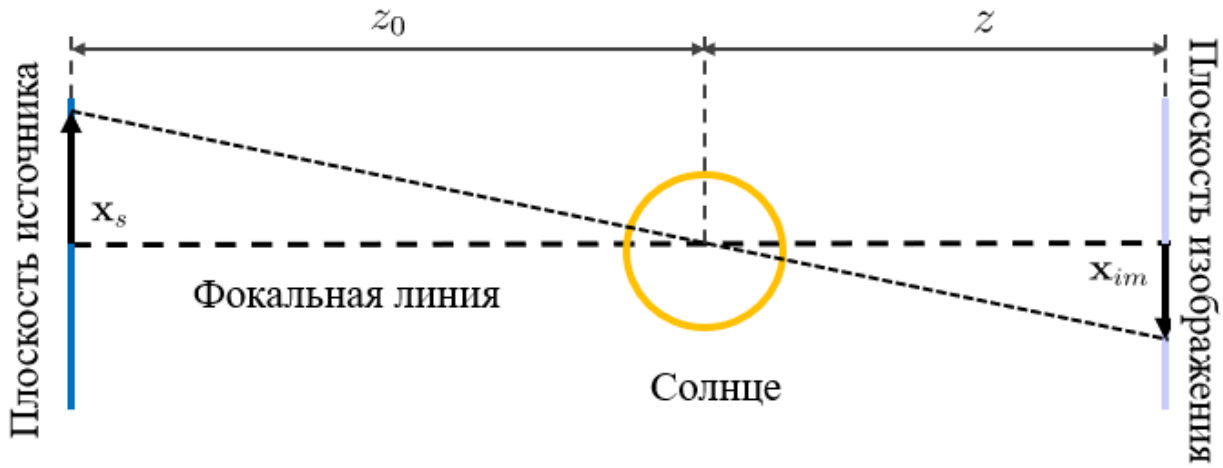

Рисунок 2. Схематическое представление процесса формирования изображения в ГФС из [1].

1.2. Общая постановка задачи движения аппарата в ГФС

Предполагается, что космический аппарат (КА) с телескопом начинает движение на расстоянии примерно 550 а.е. от Солнца, а удаление от фокальной линии не превышает 100 тыс. км в начальный момент времени. На аппарат не действуют никакие внешние силы в области ГФС (гравитация Солнца пренебрегается). Управлением считается реактивное ускорение от двигателей. Оно моделируется в импульсном приближении, ограничение на величину импульса составляет 100 м/с. Общий запас топлива задаётся через ограничение на суммарную характеристическую скорость, которая соответствует не более чем шести импульсам указанной величины. Цель алгоритма управления – переместить аппарат как можно ближе к фокальной линии ГФС за время равное 30 дням.

Проблема поиска алгоритма управления решается с помощью применения методов ОП. Более того, для упрощения задачи и для исследовательских целей, задача решается в двух постановках. В первой постановке не учитываются движения Солнца и экзопланеты: предполагается, что они зафиксированы в каких-то точках пространства. Это позволяет

упростить математическую модель и сконцентрироваться на построении базовой стратегии управления. Отброшенные в первой постановке движения Солнца и экзопланеты учитываются во второй постановке. Это ведёт к усложнению уравнений движения и повышению вычислительной нагрузки во время моделирования движения, однако позволяет найти алгоритм управления для более реалистичных условий.

1.2. Системы координат

В работе используются две системы координат. Первая система координат — инерциальная, с центром в барицентре Солнечной системы. Её оси ориентированы относительно удалённых звёзд. Параметры орбит планет, включая движение Солнца, заданы в этой системе по данным, выраженным в эклиптической системе координат на эпоху J2000.0. Вторая система координат – это так называемая фокусная система координат (ФСК), которая является неинерциальной. Она вводится следующим образом. Пусть

$\boldsymbol{R}$ – положение Солнца относительно барицентра Солнечной системы,

$\boldsymbol{r}_{bc}$ – положение относительно барицентра Солнечной системы

$\boldsymbol{r}_{bc}^{p}$ – положение экзопланеты относительно барицентра экзосистемы.

Направляющий вектор фокальной оси ГФС определяется как

$$\mathbf{n}_0 = -\frac{(\boldsymbol{r}_{bc} + \boldsymbol{r}_{bc}^{p}) - \boldsymbol{R}}{\left|(\boldsymbol{r}_{bc} + \boldsymbol{r}_{bc}^{p}) - \boldsymbol{R}\right|}, \tag{1}$$

который, в предположении $|\boldsymbol{R}| \ll |\boldsymbol{r}_{bc}|$, может быть аппроксимирован как

$$\widetilde{\boldsymbol{n}}_0 = -\mathbf{n}_{bc} - \frac{\boldsymbol{\Delta} - (\boldsymbol{\Delta}, \mathbf{n}_{bc})\mathbf{n}_{bc}}{z_0}, \tag{2}$$

где

$$\mathbf{n}_{bc} = \frac{\boldsymbol{r}_{bc}}{|\boldsymbol{r}_{bc}|}, \tag{3}$$

$$\boldsymbol{\Delta} = \boldsymbol{r}_{bc}^{p} - \boldsymbol{R}. \tag{4}$$

Также введем вектор $\boldsymbol{a}$, который является постоянным вектором из ИСК и который никогда не коллинеарен $\widetilde{\boldsymbol{n}}_0$. Такой вектор всегда можно выбрать, потому что движение фокальной линии ограничено в пространстве. С использованием введённых векторов фокусная система координат определяется следующим образом:

- начало координат совпадает с центром масс Солнца,
- ось z направлена вдоль базисного вектора

$$\mathbf{g}_3 = \frac{\widetilde{\boldsymbol{n}}_0}{|\widetilde{\boldsymbol{n}}_0|}, \tag{5}$$

- ось x направлена вдоль базисного вектора

$$\mathbf{g}_1 = \frac{\boldsymbol{a} - (\boldsymbol{a}, \boldsymbol{g}_3)\boldsymbol{g}_3}{|\boldsymbol{a} - (\boldsymbol{a}, \boldsymbol{g}_3)\boldsymbol{g}_3|}, \tag{6}$$

- ось y направлена вдоль базисного вектора

$$\mathbf{g}_2 = \mathbf{g}_3 \times \mathbf{g}_1. \tag{7}$$

При таком построении плоскость Oxy является плоскостью изображения, а ось z направлена вдоль фокальной оси ГФС (рис. 3)

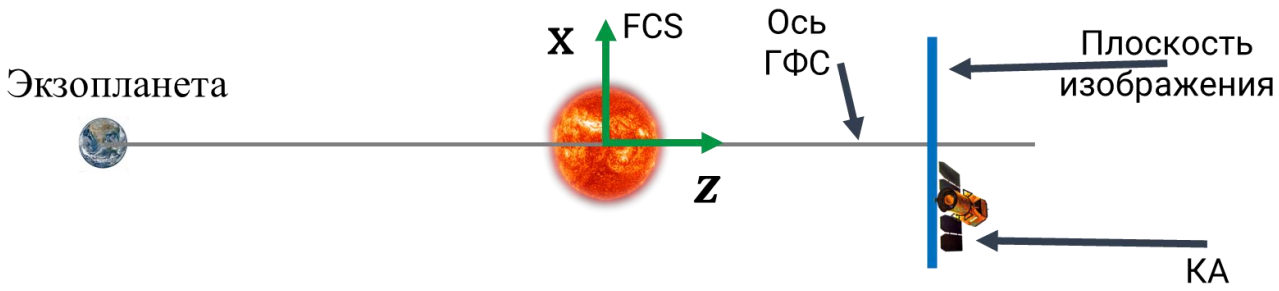

Рисунок 3. Схематическое представление ФСК.

Матрица перехода от ИСК к ФСК выглядит следующим образом:

$$\mathbf{S}(t) = \left(\mathbf{g}_1^{\text{ИСК}}(t), \mathbf{g}_2^{\text{ИСК}}(t), \mathbf{g}_3^{\text{ИСК}}(t)\right), \tag{8}$$

где $\mathbf{g}_i^{\text{иск}}$ – координатный столбец вектора $\mathbf{g}_\text{i}$, записанный в базисе ИСК. Связь координат радиуса-вектора аппарата $\boldsymbol{\tau}$ относительно ИСК, записанного в двух базисах имеет следующий вид:

$$\boldsymbol{\tau}^{\text{ИСК}} = \mathbf{S}\boldsymbol{\tau}^{\text{ФСК}}, \tag{9}$$

а связь радиусов-векторов КА в ИСК и в ФСК задаётся следующим векторным равенством:

$$\boldsymbol{r} = \boldsymbol{\tau} - \boldsymbol{R}. \tag{10}$$

1.3 Уравнения движения КА

В ИСК движение аппарата описывается вторым законом Ньютона:

$$\ddot{\boldsymbol{\tau}}^{\text{ИСК}} = \boldsymbol{a}_{\text{внеш}}^{\text{ИСК}}, \tag{11}$$

где $\boldsymbol{a}_{\text{внеш}}^{\text{ИСК}}$ – полное ускорение, вызванное реактивными силами от двигателей, записанное в ИСК. Уравнение (11) является уравнением, используемым в постановке, не учитывающей влияние движения Солнца и экзопланеты (сценарий 1). Дважды продифференцируем (10)

$$\ddot{\boldsymbol{r}}^{\text{ИСК}} = -\ddot{\boldsymbol{R}}^{\text{ИСК}} + \boldsymbol{a}_{\text{внеш}}^{\text{ИСК}}. \tag{12}$$

Чтобы записать уравнение эволюции координат в неинерциальной системе ФСК, используем уравнения (12) и (9), которое продифференцировано дважды, откуда получим:

$$\ddot{\boldsymbol{r}}^{\text{ФСК}} = -\ddot{\boldsymbol{R}}^{\text{ФСК}} + \boldsymbol{a}_{\text{внеш}}^{\text{ФСК}} - \mathbf{S}^{\top}\left(\ddot{\mathbf{S}}\boldsymbol{r}^{\text{ФСК}} + 2\dot{\mathbf{S}}\dot{\boldsymbol{r}}^{\text{ФСК}}\right). \tag{14}$$

где $\boldsymbol{r}$ – радиус-вектор, соединяющий Солнце и КА, $\boldsymbol{R}$ – радиус-вектор, соединяющий барицентр Солнечной системы и Солнце, $\mathbf{S}$ – матрица перехода от ИСК к ФСК, $\boldsymbol{u}$ – реактивное управление от двигателей.

Уравнения (14) можно записать в терминах угловых скорости и ускорения:

$$\ddot{\boldsymbol{r}}^{\text{ФСК}} = \boldsymbol{u}^{\text{ФСК}} - \ddot{\boldsymbol{R}}^{\text{ФСК}} - 2\boldsymbol{\omega}^{\text{ФСК}} \times \boldsymbol{r}^{\text{ФСК}} - \boldsymbol{\varepsilon}^{\text{ФСК}} \times \boldsymbol{r}^{\text{ФСК}} - \boldsymbol{\omega}^{\text{ФСК}} \times \left(\boldsymbol{\omega}^{\text{ФСК}} \times \boldsymbol{r}^{\text{ФСК}}\right), \quad (15)$$

где $\boldsymbol{\omega}^{\text{ФСК}}, \boldsymbol{\varepsilon}^{\text{ФСК}}$ – угловая скорость и ускорение вращения ФСК относительно ИСК, записанные в системе ФСК. Получить эти векторы можно получить из соотношений:

$$\boldsymbol{\omega}^{\text{ИСК}} = \frac{1}{2}\sum_{k=1}^{3} \mathbf{g}_k^{\text{ИСК}} \times \dot{\mathbf{g}}_k^{\text{ИСК}}, \quad (16)$$

$$\boldsymbol{\varepsilon}^{\text{ИСК}} = \frac{1}{2}\sum_{k=1}^{3} \mathbf{g}_k^{\text{ИСК}} \times \ddot{\mathbf{g}}_k^{\text{ИСК}}. \quad (17)$$

Уравнение (15) используется в постановке, учитывающей движение Солнца и экзопланеты (сценарий 2). Заметим, что $\boldsymbol{R}, \mathbf{S}$ а также их производные первого и второго порядка, как и угловая скорость и ускорение вращения ФСК относительно ИСК, являются функциями исключительно времени. Поэтому для ускорения численных расчётов эти величины были предрасчитаны на фиксированной временной сетке**,** а их значения в произвольный момент времени вычислялись с помощью сферической линейной интерполяции кватернионов для кватернионов ориентации и линейной интерполяции для угловых скоростей и ускорений.

Для иллюстрации поведения системы, описываемой уравнениями (15), представлен рисунок 4, демонстрирующий свободное движение КА в этой области (важно отметить, что это движение зависит от конкретных используемых параметров движения Солнца и экзопланеты, а также от начальных положения и скорости КА). Малые кольца вызваны влиянием движения экзопланеты, а большие – Солнца. Характерные размеры малых колец имеют порядок 100 тыс. км, больших – 1 млн км.

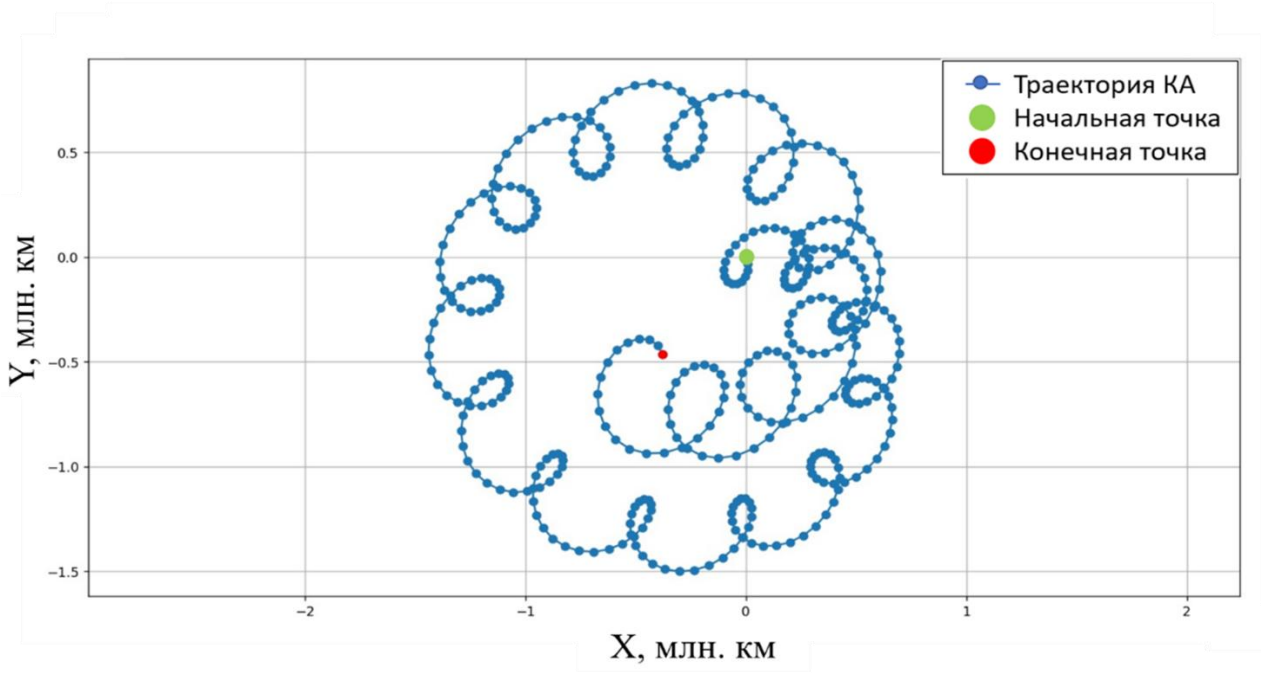

Рисунок 4. График свободного движения аппарата в области ГФС за 20 лет.

1.4 Постановка задачи оптимального управления.

Рассматривается задача управления движением космического аппарата (КА) в фокусной системе координат (ФСК). Задача рассматривается в двух постановках:

Постановка 1: Уравнения движения имеют вид

$$\ddot{\boldsymbol{\tau}}^{\text{ИСК}} = \boldsymbol{u}^{\text{ИСК}}, \tag{18}$$

где $\boldsymbol{\tau}^{\text{ИСК}}$ – радиус-вектор аппарата относительно ИСК, $\boldsymbol{u}^{\text{ИСК}}$ – реактивное управление от двигателей аппарата.

Постановка 2: Уравнения движения

$$\ddot{\boldsymbol{r}}^{\text{ФСК}} = \boldsymbol{u}^{\text{ФСК}} - \ddot{\boldsymbol{R}}^{\text{ФСК}} - 2\boldsymbol{\omega}^{\text{ФСК}} \times \boldsymbol{r}^{\text{ФСК}} - \boldsymbol{\varepsilon}^{\text{ФСК}} \times \boldsymbol{r}^{\text{ФСК}} - \boldsymbol{\omega}^{\text{ФСК}} \times \left(\boldsymbol{\omega}^{\text{ФСК}} \times \boldsymbol{r}^{\text{ФСК}}\right), \tag{19}$$

где $\boldsymbol{r}^{\text{ИСК}}$ – радиус-вектор аппарата относительно ИСК, остальные члены введены в предыдущем пункте главы.

Пусть космический аппарат изначально находится на расстоянии R_0 от фокальной линии источника и движется со скоростью $v_z^{\text{ИСК}}(0)$, параллельной фокальной линии в инерциальной системе, связанной с Солнцем; при этом вопрос предварительного выравнивания скорости не рассматривается. Тогда начальные условия для обеих постановок задаются следующим образом:

$$x^2(0) + y^2(0) \le R_0^2\,, R_0 = 100\,000 \text{ км} \tag{20}$$

$$z(0) = 550 a.e., \tag{21}$$

$$v_x^{\text{ИСК}}(0) = v_y^{\text{ИСК}}(0) = 0, \tag{22}$$

$$v_z^{\text{ИСК}}(0) = 25 \text{ а.е./ год}, \tag{23}$$

Значение было выбрано $R_0 = 100\ 000$ км исходя из оценок [7], так как это то расстояние, до которого будет виден сфокусированный свет от экзопланеты, а также можно будет осуществить навигацию. Целью алгоритма управления является подведение космического аппарата (КА) к оси фокуса на минимальную возможную дистанцию с минимальной возможной радиальной скоростью относительно оси. При этом компоненты положения и скорости вдоль оси фокуса в критерий качества не включаются. Управление осуществляется с использованием фиксированного числа импульсов n_imp = 6, равномерно распределённых по времени с постоянным интервалом Δt = 5 дней. Модуль каждого импульсного приращения скорости ограничен величиной $\Delta \mathbf{v}_{max}$ = 100 м/с.

2. Постановка задачи обучения с подкреплением

2.1. Общие сведения про ОП

В обучении с подкреплением управление динамическими системами интерпретируется как взаимодействие агента со средой: агент совершает действие, получая за это вознаграждение от среды, которая переходит в новое состояние ([8]). Управляющее воздействие формируется с помощью параметризованной стратегии, которая отображает текущее состояние (или, в общем случае, наблюдение) космического аппарата в соответствующее действие. Подбор параметров этой стратегии осуществляется таким образом, чтобы выполнялось уравнение оптимальности Беллмана и/или чтобы среднее суммарное вознаграждение агента за весь полёт по различным начальными условиям было максимальным.

Рассмотрим ситуацию, в которой агент осуществляет взаимодействие со средой в дискретные моменты времени $t = 0,1,2$ .... Будем предполагать, что состояния S_t, действия A_t и вознаграждения R_t являются случайными величинами, их конкретные значения будем обозначать строчными буквами. Обозначим через **S** множество всех возможных состояний среды, а через **A** – множество всех допустимых действий агента. Эволюция среды описывается функцией перехода, которая задаёт вероятность перехода из одного состояния в другое при выбранном действии. Для всех $s \in \mathbf{S}$, $a \in \mathbf{A}$, $s' \in \mathbf{S}$ и $t = 0,1,2$ ...эта функция имеет вид

$$p(s, a, s') = \mathbb{P}(S_{t+1} = s' | S_t = s, A_t = a). \quad (24)$$

Предположим, что вероятность перехода среды из одного состояния в другое зависит исключительно от текущего состояния и выбранного действия, и не изменяется во времени — таким образом, среда считается стационарной. Кроме того, предполагается, что переходы описываются исключительно функцией перехода $p(s, a, S')$, вероятности перехода не зависят от истории системы, будущее системы зависит только от настоящего (S_t, A_t) и не зависит от прошлых значений $S_{t-1}, S_{t-2}, \dots, A_{t-1}, A_{t-2}, \dots$ – такие системы называются марковскими. Обозначим R_t – вознаграждение в момент времени *t*. Генерацию вознаграждения описывает функция d_R, которая предполагается независящей от времени, то есть R_t есть случайная величина с распределением $d_R(S_t, A_t, S_{t+1})$.

Функцией вознаграждения будем называть функцию следующего вида

$$R(s, a) = E(R_t | S_t = s, A_t = a), \quad (25)$$

которая определяет среднее вознаграждение в момент t при данных состоянии и действии.

Стратегией является такая функция, которая для всех $s \in \mathbf{S}$, $a \in \mathbf{A}$, $s' \in \mathbf{S}$ и $t = 0,1,2$ ... есть

$$\pi(s, a) = \mathbb{P}(A_t = a | S_t = s). \quad (26)$$

Иными словами, стратегия сопоставляет каждому состоянию вероятностное распределение на множестве допустимых действий, из которого конкретное действие выбирается случайным образом в соответствии с заданными вероятностями.

Рассмотрим процесс взаимодействия агента со средой. Начальное состояние S_0 выбирается случайным образом согласно начальному распределению $d_0(s) = \mathbb{P}(S_0 = s)$, заданному для всех $s \in \mathbf{S}$. Затем, на основе текущего состояния и стратегии агент выбирает действие A_0. В результате применения действия среда, согласно функции перехода p, попадает в новое состояние S_1, а агент получается вознаграждение R_0. После этого агент выбирает следующее действие A_1, и процесс продолжается по аналогичной схеме.

В обучении с подкреплением агент направлен на максимизацию суммарного вознаграждения, получаемого от среды за определенный промежуток времени. Для этого вводится целевая функция, отражающая эффективность поведения агента

$$J(\pi) = E\left(\sum_{t=0}^{\infty} \gamma^t R_t \middle| \ \pi\right), \tag{27}$$

где введена величина $\gamma \in [0,1]$, ряд в (27) сходится с вероятностью единица при $\gamma < 1$. Все действия в (27) производятся при фиксированной стратегии π. Стратегия π^*называется оптимальной если

$$\pi^* \in \underset{\pi \epsilon \Pi}{argmax}\, J(\pi). \tag{28}$$

Вопрос же существования оптимальной стратегии разрешается следующей теоремой.

Теорема. Если $|\mathbf{S}| < \infty, |\mathbf{A}| < \infty, |R_t| < R_{\max} < \infty,\ 0 \le \gamma < 1,$ то оптимальная стратегия существует.

Опишем процесс обучения: вначале случайным образом инициализируются параметры θ стратегии $\pi(s, a; \theta)$, где θ обозначает все оптимизируемые параметры. Рассматривается общий случай управления на основе наблюдений и их истории. Затем проводится оценка стратегии с помощью серии испытаний (эпизодов) Монте–Карло. В каждом эпизоде система инициализируется начальными условиями $(t_0, \boldsymbol{x}_0)$, выбранными из заданного распределения $\boldsymbol{D}_0$ на области начальных состояний Ω_0. Состоянию сопоставляется наблюдение $\mathbf{o}_0$, которое подаётся на вход стратегии. Стратегия вырабатывает действие $\mathbf{a}_0$ и обновлённую историю наблюдений или скрытое состояние $\mathbf{h}_0$. Пара $(\boldsymbol{x}_0, \mathbf{u}_0)$ используется для расчёта нового состояния системы, значения вознаграждения и флага завершения эпизода (конец эпизода определяется достижением заранее заданного количества шагов взаимодействия агента со средой или выполнением критерия останова). Если эпизод не завершён, цикл повторяется для нового состояния. После завершения эпизода процесс начинается заново с другим начальными условиями.

Многократное повторение этого цикла приводит к накоплению набора данных: состояний s, действий a, наблюдений o, вознаграждений r и флагов завершения d. Эти данные передаются в алгоритм обучения с подкреплением (в работе был применен алгоритм PPO), который обновляет параметры стратегии. Затем серия испытаний повторяется уже для обновлённой стратегии.

2.2. Описание алгоритма PPO

Алгоритм Proximal Policy Optimization —это метод обучения с подкреплением, направленный на предотвращение чрезмерно резкого изменения стратегии в процессе обновления параметров $\pi(s, a, \theta)$. Для этого используется модифицированный целевой функционал, называемый обрезанным функционалом, обозначаемый как $J^{clip}(\theta)$. Он определяется следующим образом:

$$J^{clip}(\theta) = \mathbf{E}\, min\left(\rho(\theta) \cdot A^{\theta_k}(s,a), \qquad clip(\rho(\theta), 1-\epsilon, 1+\epsilon) \cdot A^{\theta_k}(s,a)\right), \tag{29}$$

где $\rho(\theta) = \pi(s, a, \theta)\,/\pi(s, a, \theta_k)$, ϵ – гиперпараметр, ограничивающий допустимую величину отклонения вероятности совершения того или иного действия при обновлении стратегии, $A^{\theta_k}(s, a)$ – функция преимущества, вычисленная для стратегии $\pi(s, a, \theta_k)$, $\mathbf{E}[\cdot]$ — математическое ожидание, то есть усреднение по всем наблюдаемым парам состояний и действий из выборки. Под $A(s, a)$ подразумевается функция преимущества, которая равна разности суммарного вознаграждения после действия a в состоянии s и среднего вознаграждения при действии в соответствии со стратегией π

$$A(s, a) = q^{\pi}(s, a) - v^{\pi}(s). \tag{30}$$

Таким образом, если $A(s, a) > 0$, то действие a приводит к вознаграждению большему, чем если бы агент действовал в соответствии со стратегией π (в среднем), а если $A(s, a) < 0$, то, наоборот, действие а приводит к вознаграждению меньшему, чем среднее вознаграждение при действиях в соответствии со стратегией π.

Функция $clip$ здесь играет важную роль: она ограничивает значение $\rho(\theta)$ в пределах от $1-\epsilon$ до $1+\epsilon$, что предотвращает чрезмерные обновления стратегии, особенно в случаях, когда преимущество $A(s, a)$ имеет большой модуль. Это обеспечивает устойчивость обучения, так как обновления происходят в ограниченном доверительном интервале относительно предыдущей стратегии.

Схематично алгоритм PPO представлен на рисунке 5: в начале происходит инициализация весов функции стратегии $\pi(s, a, \theta)$ и ценности состояния $v^{\pi}(s, \omega)$, которая тоже будет обучаться в процессе. Здесь ω – обучаемые параметры. Далее запускается основной цикл алгоритма. В каждом цикле агент взаимодействует со средой на протяжении заданного количества шагов (один эпизод может состоять из нескольких шагов). В процессе

этих шагов собирается траектория – история взаимодействий агента со средой. Симулируется некоторое количество эпизодов, рассчитываются оценки вознаграждения и функции преимущества, а затем происходит обновление весов θ, ω.

Алгоритм Proximal policy optimization (PPO).

Инициализация начальных параметров функции стратегии θ_0 и функции ценности ω_0;

Для каждого $k = 0,1,2, \ldots$ выполняется

Сбор истории взаимодействия агента со средой $D_k = \{\tau_i\}$ с использованием стратегии $\pi(\cdot,\cdot,\theta_k)$;

Расчет наград $\hat{R}_t = \sum_{k=0}^{T-t} R_{t+k}\gamma^k$;

Расчет оценки функции преимущества $\hat{A}_t$ (используя любой метод оценки преимущества) на основе текущего значения функции ценности $v(\cdot,\omega_k)$;

Обновление весов стратегии:

$$\theta_{k+1} = arg\max_{\theta} \frac{1}{|D_k|T} \sum_{\tau_\epsilon D_k} \sum_{t=0}^{T} min\left(\rho(\theta)\hat{A}_t, clip(\rho(\theta), 1-\epsilon, 1+\epsilon)\cdot \hat{A}_t\right),$$

где $\rho(\theta) = \pi(s,a,\theta)\,/\pi(s,a,\theta_k)$;

Обновление весов функции ценности

$$\omega_{k+1} = arg\min_{\omega} \frac{1}{|D_k|T} \sum_{\tau_\epsilon D_k} \sum_{t=0}^{T} \left(v(s_t,\omega) - \hat{R}_t\right)^2;$$

Конец

Рисунок 5 – Описание алгоритма PPO из [8].

2.3. Сведение задачи оптимального управления к задаче ОП

Преобразование задачи выполним в соответствии с методикой, описанной в [6]. Для того, чтобы свести механическую задачу оптимального управления к задаче обучения с подкреплением необходимо выполнить следующие шаги:

В первую очередь стоит связать состояние **X** (фазовый вектор механической системы) с состоянием ***s*** (определяемым в обучении с подкреплением). В данной работе они совпадают **X** = ***s***.

Далее, на множестве начальных состояний Ω_0 следует определить начальное распределение состояний $d_0(s) = P(S_0 = s)\ \ \forall s \in \Omega_0$. В этой работе было использовано нормальное распределение вероятностей на множестве

$$\Omega_0 = \{x^2(0) + y^2(0) \le R_0^2, z(0) = 550\ a.e., v_x = v_y = 0, v_z = 25 \text{ а. е./ год}\}.$$

Начальное состояние выбиралось в соответствии с постановкой задачи (20)-(23) и этим распределением.

После этого необходимо выбрать дискретное отображение (шаг по времени), которое будет сопоставлять текущим значениям времени, фазового вектора и

управляющего воздействия новое значение времени, новый фазовый вектор и флаг завершения эпизода. В начале шага управляющее воздействие применяется как мгновенный импульс, то есть прибавляется к текущей скорости аппарата. Затем отображение реализуется через численное интегрирование системы уравнений (18) для первой постановки задачи и уравнений (19) — для второй. В результате получаем новое состояние системы в следующий момент времени, то есть фазовый вектор, уже с учётом импульсного воздействия. Дополнительно формируется флаг завершения эпизода — он активируется, если достигнуто заранее заданное число шагов.

Также необходимо определить функцию вознаграждения r, которая сопоставляет численное значение (награду) элементам перехода системы. В общем случае она зависит от текущего времени t, текущего фазового состояния $\mathbf{x}_t$, управляющего воздействия $\mathbf{a}_t$, следующего времени $t+1$, следующего фазового состояния $\mathbf{x}_{t+1}$, а также флага завершения эпизода d_{t+1}. В работе была использована следующая функция вознаграждения:

$$r_k = \rho(\mathbf{x}_k) - \rho(\mathbf{x}_{k+1}), \rho(\mathbf{x}_k) = \sqrt{{x_k}^2 + {y_k}^2} + \alpha\sqrt{v^2_{x_k} + v^2_{y_k}}, \tag{31}$$

где α – весовой коэффициент, равный единице в безразмерной системе единиц.

Времени и фазовому вектору необходимо сопоставить наблюдение, то есть определить модель восприятия. В постановке 1 наблюдениями являлся фазовый вектор. В постановке 2 наблюдения состояли из двух предыдущих фазовых векторов и приращений скорости, текущего фазового вектора, а также разности $\mathrm{d_k}$ текущего фазового вектора и предсказания фазового вектора в модели, не учитывающей движение Солнца и экзопланеты: $\mathbf{o} = [\mathbf{x}_{k-2}, \mathbf{x}_{k-1}, \mathbf{x}_k, \Delta\boldsymbol{v}_{k-1}, \Delta\boldsymbol{v}_k, \mathrm{d_k}]$.

Следует выбрать модель управления, которая сопоставляет наблюдению вектор управления. В качестве модели управления рассматривается нейросетевая модель с одним скрытым слоем

$$\mathbf{a}(\mathbf{o}) = \mathbf{A}_2^{\pi}\,\mathrm{th}(\mathbf{A}_1^{\pi}\mathbf{o} + \mathbf{b}_1^{\pi}) + \mathbf{b}_2^{\pi}, \qquad \mathbf{u} = \mathbf{a}\cdot\min(1, \Delta v_{\max}/|\mathbf{a}|), \tag{32}$$

где $\mathbf{a}$ **–** действие агента, $\mathbf{u}$ – управление, $\mathbf{o}$ – наблюдение, $\mathbf{A}_1^{\pi}, \mathbf{A}_2^{\pi}, \mathbf{b}_1^{\pi}, \mathbf{b}_2^{\pi}$ - оптимизируемые параметры стратегии, которые входят в совокупность параметров θ. Функция ценности аппроксимируется функцией

$$v(\mathbf{o}) = \mathbf{a}_2^{vT}\,\mathrm{th}(\mathbf{A}_1^{v}\mathbf{o} + \mathbf{b}_1^{v}) + \mathbf{b}_2^{v}, \tag{33}$$

где параметры $\mathbf{a}_2^{vT}, \mathbf{A}_1^{v}, \mathbf{b}_1^{v}, \mathbf{b}_2^{v}$ составляют вектор параметров ω.

После осуществления всех шести шагов механическая задача трансформируется в задачу машинного обучения с подкреплением.

3. Результаты

3.1. Параметры обучения

Для решения поставленной задачи областью начальных условий являлся круг радиуса $R_0 = 100000$ км, расположенный в плоскости изображения (плоскость XY). Движение начинается на расстоянии z = 550 а.е. от Солнца, а экзопланета расположена на расстоянии z_0 = 10 св. лет от Солнца. Каждый эпизод продолжается 30 дней. Аппарату необходимо максимально приблизиться к фокальной линии с использованием импульсов, количество которых равно `n_imp` = 6, а интервал между импульсами Δt = 5 суток. Также импульсы ограничены по модулю величиной $\Delta \mathrm{v}_{max}$ = 0.1 км/сек. В ходе обучения и вычислений положение и скорость аппарата задаются в безразмерной системе единиц, где за единицу расстояния принимается 100 000 км, а за единицу времени — 100 000 сек. Функцией вознаграждения для обоих сценариев являлась функция

$$r_k = \rho(\mathbf{x}_k) - \rho(\mathbf{x}_{k+1}), \rho(\mathbf{x}) = \sqrt{x^2 + y^2} + \alpha\sqrt{v_x^2 + v_y^2}, \quad (34)$$

где α – весовой коэффициент, равный единице в безразмерной системе единиц.

Обучение производилось с помощью алгоритма PPO с использованием библиотек stable-baselines3 и kiam-rl. Количество нейронов на скрытом слое выбрано равным n_1 = 6 для первой постановки задачи (без движения фокальной линии) и n_2 = 32 для второй постановки задачи. Для аппроксимации среднего значения J функционала был установлен объём выборки (число шагов, задаётся параметром n_steps) равный 10 000. Число итераций градиентного спуска для обновления весов нейросетевых моделей (параметр n_epochs) было задано равным 30, а значение скорости обучения (learning_rate) — 0.001. Обучение производилось на центральном процессоре и завершалось по достижении 1.5 миллиона шагов (параметр total_timesteps). Параметры были взяты из [6]. Дополнительно, в процессе обучения использовался параметр `std`, управляющий дисперсией стохастической политики. Он влияет на то, насколько «разнообразные» действия агент будет пробовать на начальных этапах. На практике это помогает избежать сходимости к неоптимальным стратегиям. Значение `std` подбиралось эмпирически и было выбрано значение равное -3 для обоих сценариев.

3.2. Результаты в постановке без движения фокальной линии

В данном сценарии за наблюдения принимался фазовый вектор аппарата $\mathbf{o} = \mathbf{X}$, т.е. предполагалось, что известен радиус вектор и скорость аппарата в текущий момент времени. В ходе обучения оценка среднего суммарного вознаграждения выросла со значения -2.08 на первых итерациях до 0.664 на последних итерациях. Отметим, что для теоретически оптимальной стратегии, полностью устраняющей отклонения по положению

и скорости относительно фокальной линии и, тем самым, максимизирующей ожидаемое суммарное вознаграждение, среднее значение этого вознаграждения составляет

$$\int_0^{R_0} \int_0^{2\pi} \frac{r^2}{\pi} dr\, d\varphi = \int_0^1 2r^2\, dr = \frac{2}{3} \approx 0.6667. \quad (35)$$

Это говорит о том, что найденное управление близко к оптимальному.

Промах по положению и скорости в конце эпизода можно объяснить тем, что нейросеть имеет ограниченные возможности, а процесс оптимизации сходится только к локальному минимуму. При этом сам функционал тоже считается приближённо — на основе серии запусков по методу Монте-Карло.

Качество полученного управления было проверено на 5000 испытаниях Монте-Карло, где начальные условия выбирались также, как и при обучении (равномерно из круга радиусом R_0). В таблице 1 показаны оценки промаха по положению и скорости в конце эпизода, а также суммарные затраты характеристической скорости.

	q0	q0.25	q0.5	q0.75	q1.0	μ
Δr_f, км	84.95	1887.23	2631.28	3181.86	6561.37	2613.21
Δv_f, м/с	0.28	13.07	19.56	24.87	37.08	18.91
u, м/с	68.92	82.36	90.69	96.64	107.53	89.30

Таблица 1. Результаты обучения для сценария 1: квантили и средние значения распределений промаха по положению Δr_f и скорости Δv_f мимо фокальной линии и суммарные затраты за эпизод характеристической скорости u.

На рисунке 6 показаны начальные и конечные расстояния до фокальной линии, на рисунке 7 приведены те же начальные и конечные расстояния, но без линии равенства. Для наглядности приведена линия равенства начальных и конечных расстояний. На графике все синие точки находятся под линией, что говорит о том, что конечные расстояния сильно меньше начальных. Это, в свою очередь, означает, что КА под действием найденного управления приблизился к фокальной линии вне зависимости от начальных условий (в пределах множества Ω_0).

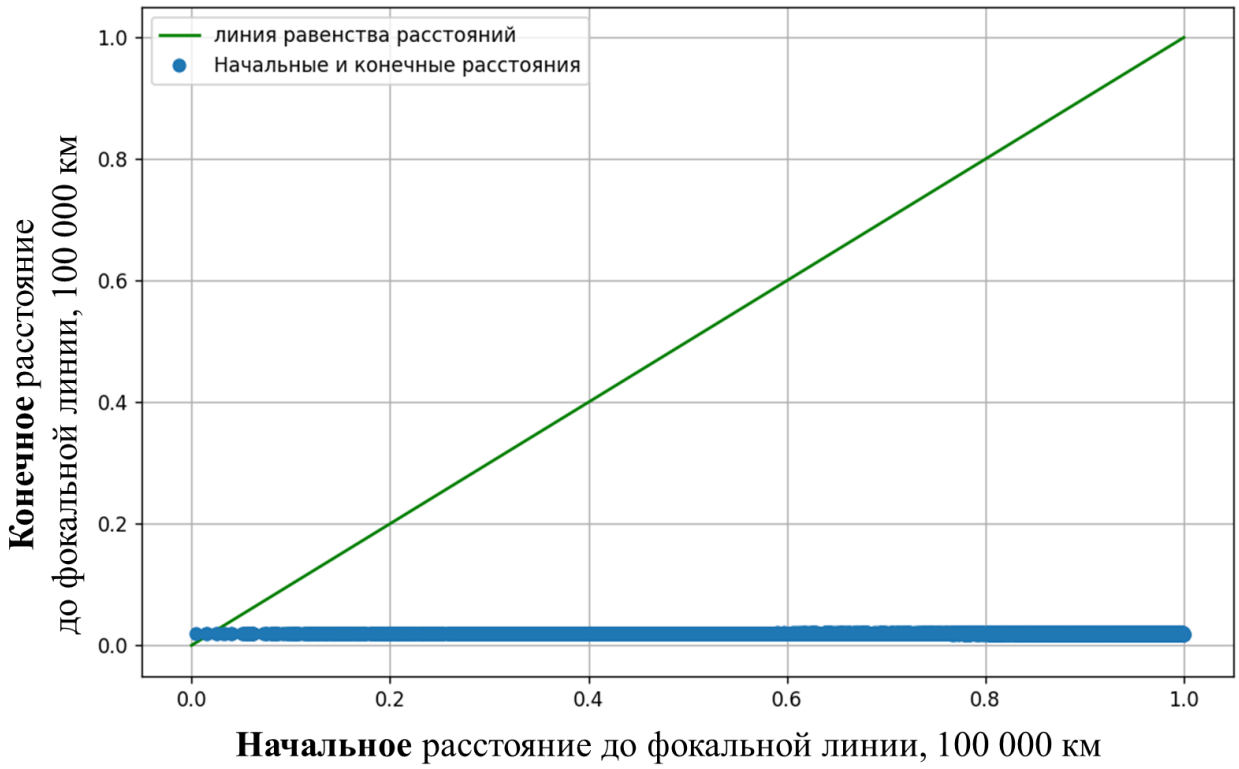

Рисунок 6. График зависимости конечных расстояний аппарата от начальных, полученных в серии испытаний Монте-Карло для постановки 1; зеленым цветом обозначена линия равенства расстояний. Подробный график поведения при малых расстояниях показаны на рисунке 7.

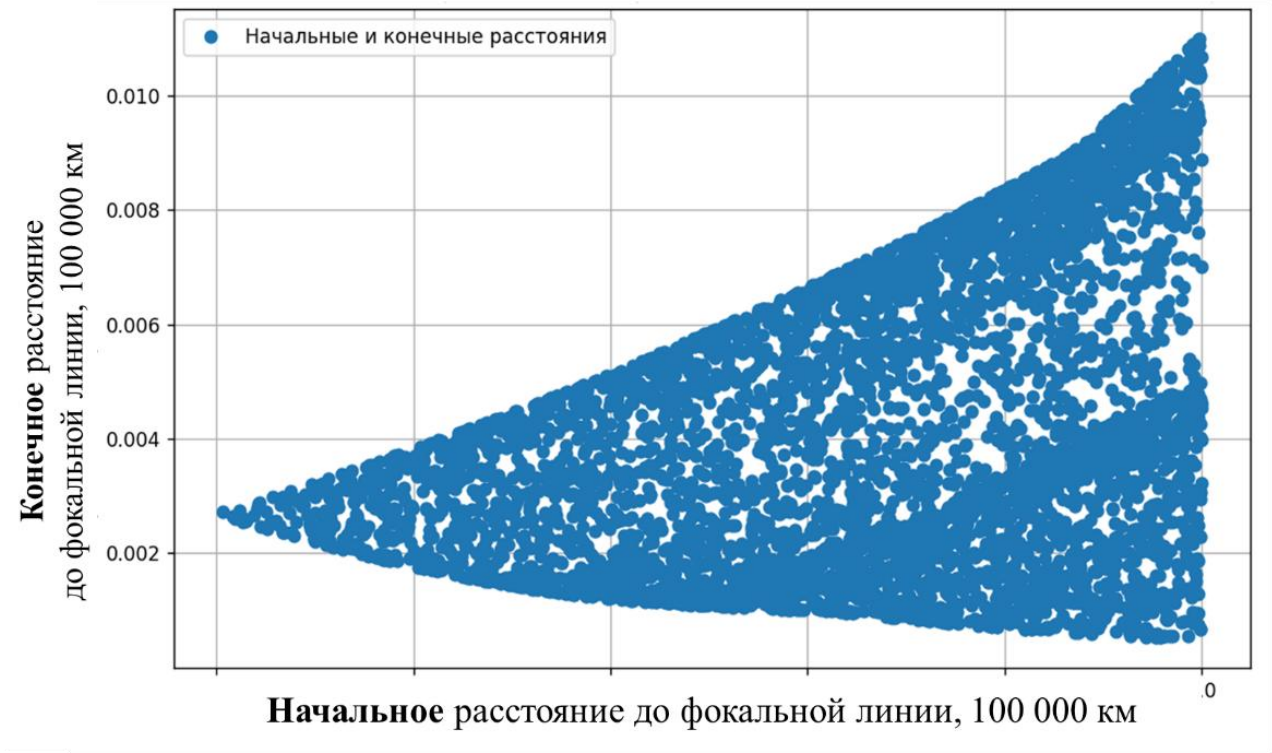

Рисунок 7. График зависимости конечных расстояний аппарата от начальных, полученных в серии испытаний Монте-Карло для постановки 1. Увеличение графика из рисунка 6. Демонстрирует детали распределения при малых начальных расстояниях.

3.3. Результаты применения модели из постановки 1 к постановке 2

Закон управления, полученный в постановке без движения фокальной линии, был применен во второй постановке, учитывающей движение Солнца и экзопланеты. Результат применения описывается графиком на рисунке 8. Исходя из этого графика, можно сделать вывод: во всех проведённых испытаниях аппарат удалялся от фокусной линии. Это означает, что полученный закон управления неприменим ко второй постановке и требуется обучение новой модели, которая соответствует постановке. На рисунке 9 приведен тот же график, но без линии равенства расстояний и в большем масштабе.

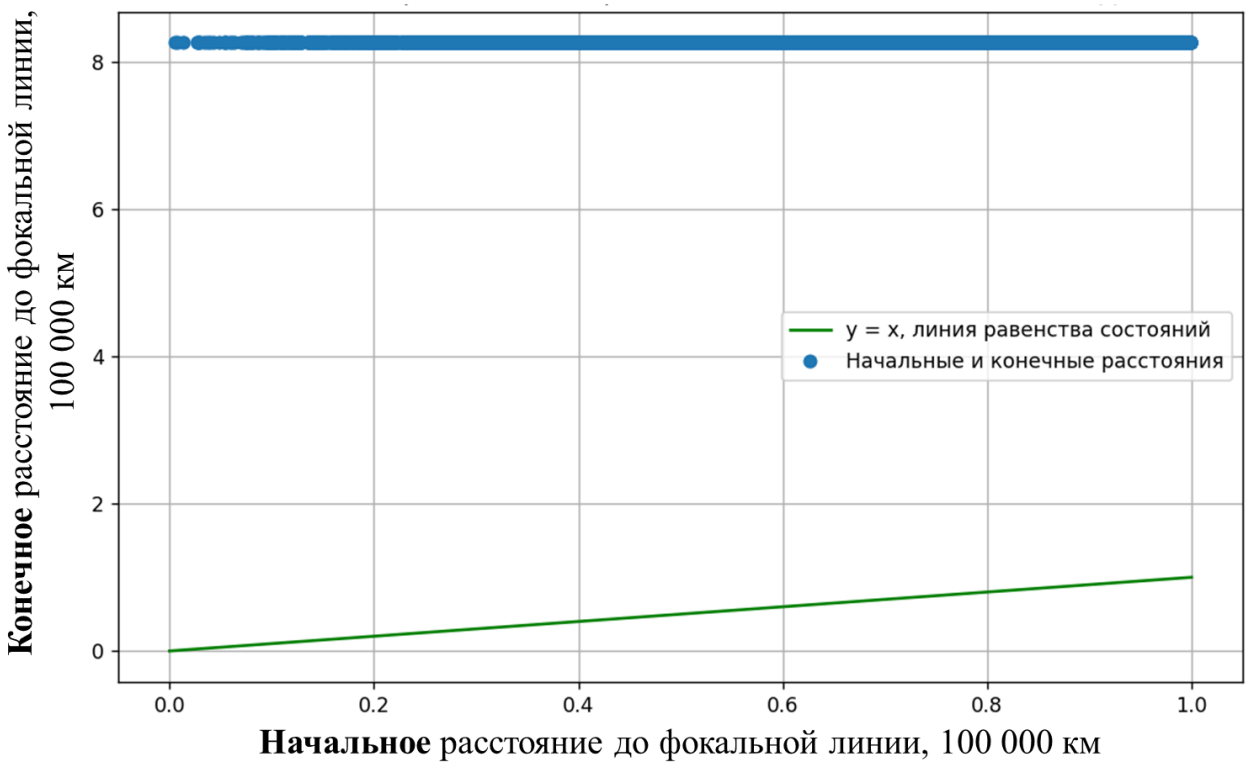

Рисунок 8. График зависимости начальных расстояний аппарата от конечных, полученных в серии испытаний Монте-Карло для сценария 2 с использованием модели, обученной в сценарии 1; зеленым цветом обозначена линия равенства расстояний. Подробный график поведения при малых расстояниях показаны на рисунке 9.

Рисунок 9. График зависимости конечных расстояний аппарата от начальных, полученных в серии испытаний Монте-Карло для постановки 2 с использованием модели, обученной в сценарии 1. Увеличение графика из рисунка 8. Демонстрирует детали распределения при малых начальных расстояниях.

3.4. Результаты в постановке 2

Переходя к модели, в которой учтено движение Солнца и экзопланеты, была использована ФСК, в уравнениях движения теперь имеют вид (14). В результате экспериментов стало ясно, что получить качественный закон управления, основывающийся на наблюдениях $\mathbf{o} = \mathbf{X}$, не получится, так как уравнения движения неавтономны и необходимы данные о текущем моменте времени. Строго говоря, нет нужды в самом времени, необходима лишь оценка неинерциальных ускорений в момент выдачи управляющего воздействия. Ввиду этого, в состав наблюдений были включены фазовые векторы и управляющие воздействия с двух предыдущих шагов по времени, а также разница наблюдаемых фазовых векторов с предсказанием фазовых векторов согласно модели (11). Поэтому в качестве наблюдения на k-ом шаге было принято использовать следующие данные:

- текущий фазовый вектор $\mathbf{X}_k \in \mathbb{R}^6$ (положение и скорость),
- два предыдущих фазовых вектора $\mathbf{X}_{k-1},\ \mathbf{X}_{k-2} \in \mathbb{R}^6$,
- два последних управляющих импульса $\Delta\boldsymbol{v}_{k-1}, \Delta\boldsymbol{v}_{k-2} \in \mathbb{R}^3$,

- разности между реальным $\mathbf{X}_k$ и предсказанным $\mathbf{X}_k^{pred}$ по модели (11), в которой не учтено движение Солнца и экзопланеты, состояниями в моменты времени t_k и t_{k-1}: $\mathrm{d_k} = C \cdot \left(\mathbf{X}_k - \mathbf{X}_k^{pred}\right)$, C – постоянная.

Такой состав наблюдений позволяет агенту учитывать краткосрочную историю движения и аппроксимировать неинерциальные ускорения. Для задания вектора наблюдений для нулевого и первого шага по времени использовалось численное интегрирование в обратном времени.

В ходе обучения для данной постановки задачи оценка среднего суммарного вознаграждения выросла с -1.748 на первых итерациях до 0.5888 на последних итерациях. Такие результаты показывают, что управление, полученное в этой постановке задачи также близко к оптимальному. Аналогично пункту 1 этой главы качество полученного управления было проверено на 5000 испытаниях Монте-Карло, результаты приведены в таблице 2.

	q0	q0.25	q0.5	q0.75	q1.0	μ
Δr_f, км	5.47	196.86	265.98	374.28	1259.22	299.35
Δv_f, м/с	0.01	0.50	0.81	1.20	3.51	0.90
u, м/с	17.75	186.82	216.20	232.20	232.72	206.45

Таблица 2. Результаты обучения для сценария 2: квантили и средние значения распределений промаха по положению $\Delta\mathrm{r_f}$ и скорости $\Delta\mathrm{v_f}$ мимо фокальной линии и суммарные затраты за эпизод характеристической скорости u.

На рисунке 10 представлены начальные и конечные расстояния до фокальной линии для каждого эпизода. Также для сравнения добавлена линия равенства расстояний. Расположение всех синих точек ниже этой линии свидетельствует о том, что конечные расстояния значительно меньше начальных. Это указывает на успешное приближение аппарата к фокальной линии независимо от его исходного положения. На рисунке 11 приведен тот же график, что и на рисунке 10, но без линии равенства расстояний и в большем масштабе.

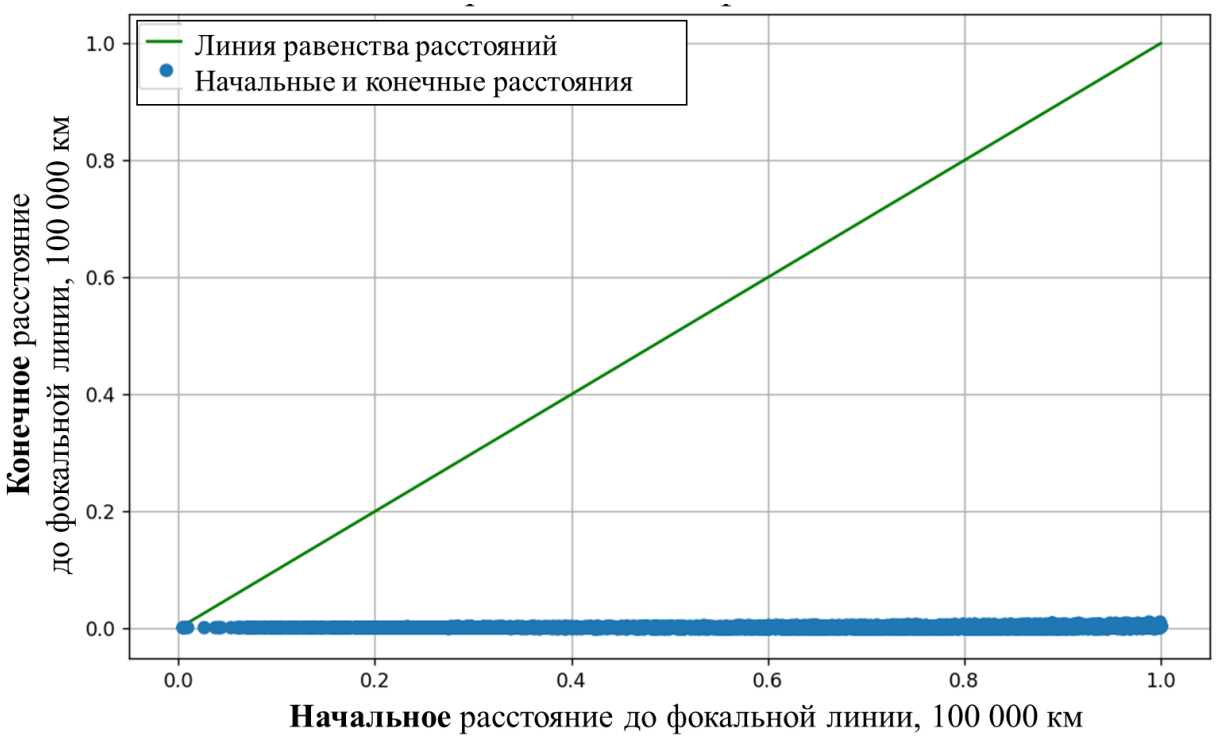

Рисунок 10. График зависимости конечных расстояний аппарата от начальных, полученных в серии испытаний Монте-Карло для постановки 2; зеленым цветом обозначена линия равенства расстояний. Подробный график поведения при малых расстояниях показаны на рисунке 11.

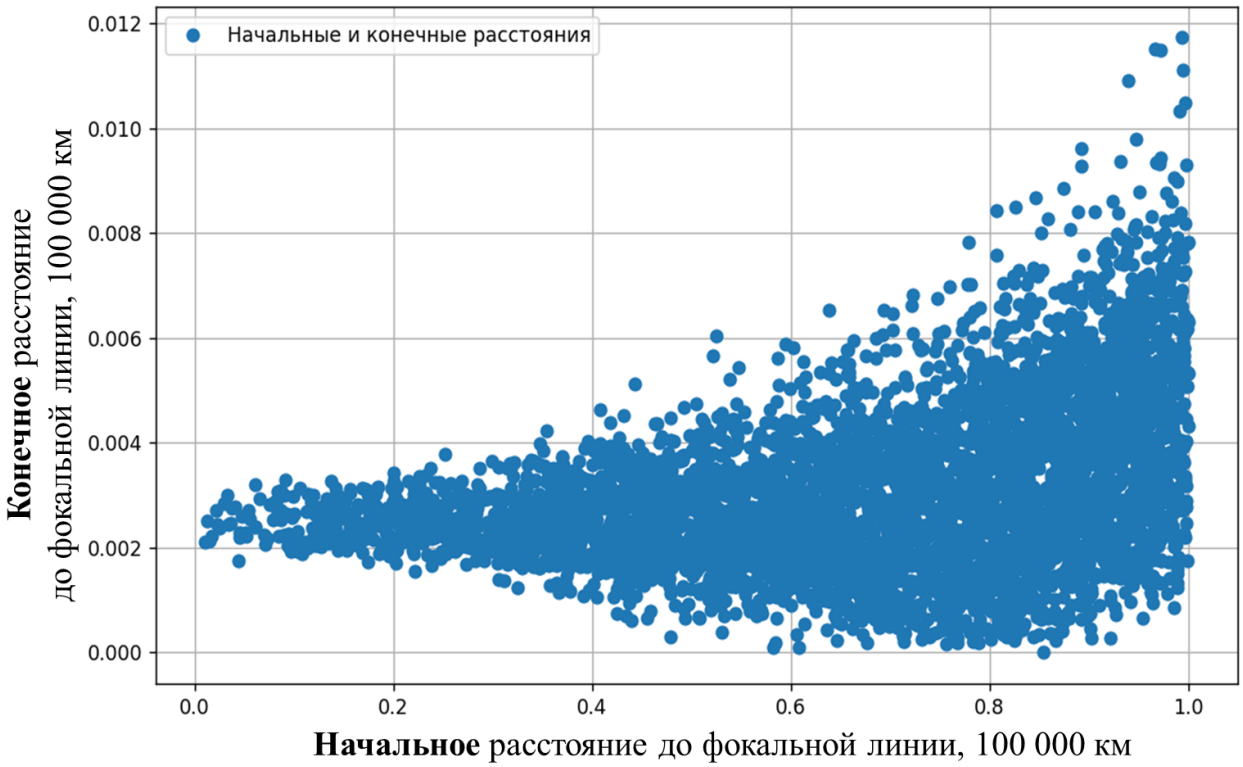

Рисунок 11. График зависимости конечных расстояний аппарата от начальных, полученных в серии испытаний Монте-Карло для постановки 2.

Увеличение графика из рисунка 19. Демонстрирует детали распределения при малых начальных расстояниях.

Заключение

В ходе работы исследовалось применение методов обучения с подкреплением для управления космическим аппаратом в области гравитационного фокуса Солнца (ГФС). Основной задачей было разработать автономное управление с обратной связью по состоянию для начального этапа миссии, когда необходимо приблизить аппарат к оси ГФС, но при этом не требуется высокоточное позиционирование.

Задача была решена в двух постановках. В первой использовалась упрощённая модель движения, игнорирующая движения Солнца и экзопланеты. В этой постановке агент успешно обучился управлять аппаратом: среднее вознаграждение почти достигло теоретического максимума, а анализ поведения показал стабильное приближение аппарата к оси ГФС. Также был получен промах по положению, который варьируется в пределах от 84.95 км до 6561.37 км, а по скорости от 0.28 м/сек и до 37.08 м/сек.

Во второй постановке применялась более сложная модель с учётом движений Солнца и экзопланеты. Попытка напрямую перенести стратегию из первой постановки оказалась неэффективной, что указало на необходимость адаптации стратегии управления под новую модель динамики. Для этого был расширен набор наблюдений агента, усложнена модель управления, и обучена новая модель управления. Это позволило значительно улучшить качество управления и обеспечить приближение аппарата к оси ГФС во второй постановке. В этой постановке промах по положению находился в пределах от 5.47 км до 1259.28 км, а по скорости от 0.01 м/сек до 3.51 м/сек.

Таким образом, проведённое исследование подтвердило, что обучение с подкреплением является перспективным подходом для решения задачи автономного управления космическим аппаратом в условиях гравитационного фокуса Солнца. При этом для успешного применения метода важно точно учитывать особенности динамики системы и правильно формировать наблюдения агента, особенно при переходе от упрощённых моделей к более реалистичным.

В будущем планируется расширение работы путем решения задачи навигации по кольцам Эйнштейна, сформированным гравитационной линзой, с целью повышения точности позиционирования аппарата в области гравитационного фокуса и обеспечения устойчивого слежения за изображением экзопланеты.

Список литературы

[1] Perepukhov D., Shirobokov M., Korneev K. On the dynamics and control of a spacecraft observing exoplanets via the Solar Gravitational Lens // *Proc. Int. Astronaut. Congr.* 2023. Baku, Azerbaijan.

[2] Широбоков М.Г. Методика построения управления космическими аппаратами с использованием методов обучения с подкреплением // Космические исследования. — 2024. — Т. 62, № 5. — С. 498–515. https://doi.org/10.31857/S0023420624050082

[3] Guzzetti D. Reinforcement learning and topology of orbit manifolds for stationkeeping of unstable symmetric periodic orbits // *AAS/AIAA Astrodynamics Specialist Conference*. 2019. 13–17 January, Ka'anapali, Maui, HI, USA.

[4] Holt H., Baresi N. Towards Optimal Lyapunov Controllers for Low-thrust Lunar transfers via Reinforcement Learning // *AAS Astrodynamics Specialist Conference*. 2019. 13–17 January, Ka'anapali, Maui, HI, USA.

[5] Das-Stuart A., Howell K. Contingency planning in complex dynamical environments via heuristically accelerated reinforcement learning // *AAS/AIAA Astrodynamics Specialist Conference*. 2019. 13–17 January, Ka'anapali, Maui, HI, USA.

[6] Широбоков М.Г., Корнеев К.Р., Перепухов Д.Г. Автономное управление космическим аппаратом в области фокуса гравитационной линзы Солнца на основе методов машинного обучения с подкреплением // Космические исследования. — 2025. — Т. 63, № 2. — С. 204–220. https://doi.org/10.31857/S0023420625020072

[7] Helvajian H., Rosenthal A., Poklemba J. и др. Mission Architecture to Reach and Operate at the Focal Region of the Solar Gravitational Lens // Journal of Spacecraft and Rockets. — 2023. — Т. 60, № 3. — С. 829–847. https://doi.org/10.2514/1.A35493 .

[8] Широбоков М.Г. Искусственный интеллект в научных исследованиях: лекции / Факультет управления и прикладной математики МФТИ. – Долгопрудный: МФТИ, 2024. – 118 с.