КОСМИЧЕСКИЕ ИССЛЕДОВАНИЯ, 2025, том 63, № 2, с. 204–220

УДК 629.78,004.852

АВТОНОМНОЕ УПРАВЛЕНИЕ КОСМИЧЕСКИМ АППАРАТОМ В ОБЛАСТИ ФОКУСА ГРАВИТАЦИОННОЙ ЛИНЗЫ СОЛНЦА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

© 2025 г. М. Г. Широбоков*, К. Р. Корнеев, Д. Г. Перепухов

Институт прикладной математики им. М.В. Келдыша РАН, Москва, Россия

**e-mail: shirobokov@keldysh.ru*

Поступила в редакцию 11.12.2023 г.
После доработки 26.02.2024 г.
Принята к публикации 26.03.2024 г.

Формулируется задача автономного управления поступательным движением космического аппарата в окрестности фокуса гравитационной линзы Солнца. Поставленная задача решается методом машинного обучения с подкреплением с использованием современных стохастических численных методов. Исследуются затраты характеристической скорости на нацеливание на фокусную линию удаленного протяженного источника, финальная точность нацеливания и качество работы функции управления. Результаты исследования приводятся для различных форм состояния и наблюдения: 1) положение и скорость; 2) зашумленные положение и скорость; 3) изображение кольца Эйнштейна. Сравнивается эффективность работы стратегий управления при использовании рекуррентных слоев и полносвязных слоев с входом в виде стека измерений. Также рассматривается обучение моделей управления с учетом ошибок исполнения маневров.

DOI: 10.31857/S0023420625020072, **EDN:** GNCBWV

1. ВВЕДЕНИЕ

Человечество стремится расширять границы наблюдаемого мира. Прошлый век был полон исследований околоземного, окололунного пространства и всех крупных небесных тел Солнечной системы. Начало XXI в. ознаменовалось изучением малых тел Солнечной системы, развертыванием крупных обсерваторий для построения карт Вселенной в различных диапазонах энергий. Следующим логичным шагом в изучении окружающего мира станет исследование пространства за пределами внутренней области Солнечной системы (более 100 а.е. от Солнца) и, что особенно интересно, наблюдение далеких экзопланет, поиск на них признаков жизни. Такие проекты имеют высокую ценность для всего человечества и порождают множество новых задач, которые не встречались для миссий в околоземном пространстве и в рамках внутренней области Солнечной системы.

На сегодняшний день только пять миссий были направлены к внешним областям Солнечной системы (*Pioneer 10*, *11*, *Voyager 1*, *2*, *New Horizons*), но ожидается, что в скором времени это число увеличится. Это связано, с одной стороны, с совершенствованием космических технологий (повышением живучести электронных компонентов, более надежными и экономичными двигателями малой тяги и солнечными парусами), а с другой – с появлением новых результатов физико-математических исследований, касающихся внешней области Солнечной системы. Например, большую ценность представляют гелиофизические исследования во внешних областях Солнечной системы. Здесь можно привести пример разрабатываемой в Лаборатории прикладной физики Университета Джонса Хопкинса (Johns Hopkins University Applied Physics Laboratory) миссии *Interstellar Probe* [1], аппарат которой будет снабжен физическими инструментами, отсутствующими на *Pioneer* и *Voyager*, для исследования гелиосферы и межзвездного пространства. Другой пример – проект Лаборатории реактивного движения (NASA's Jet Propulsion Laboratory),

предполагающий использование солнечной гравитационной линзы [2–4] для повышения эффективности телескопов и получения изображений экзопланет с большей точностью, чем позволяют обычные телескопы на околоземной орбите [5]. Проведено множество исследований по изучению оптических характеристик солнечной гравитационной линзы [6–8], разработаны методы обработки видимых изображений удаленных объектов (колец Эйнштейна) [9–11]. Показано, что телескопы, расположенные в фокусе солнечной гравитационной линзы, начинающемся на расстоянии 550 а.е. от Солнца, способны предоставить изображения экзопланет с разрешением порядка 10 км – точностью, недостижимой при использовании стандартных телескопических средств в околоземном пространстве [5]. Предварительный анализ такой миссии по наблюдению экзопланет включает оценку ее осуществимости [12, 13], однако не прорабатывает вопросы, связанные с динамикой космического аппарата и синтезом системы навигации, наведения и управления.

При подготовке какой-либо миссии во внешние области Солнечной системы возникает потребность в создании автономной системы управления, навигации и наведения, так как подобный проект рассчитан на длительное время, расстояния между Землей и аппаратом оказываются велики, сигналы распространяются с существенными задержками, Солнце может затмевать Землю, а в работе аппаратуры могут возникнуть различные проблемы. Требуется система управления и навигации, которая могла бы помочь в осуществлении такой миссии.

В настоящее время стремительно развивается и вызывает большой интерес раздел приближенного динамического программирования, известный как *машинное обучение с подкреплением* [14–16]. Обучение с подкреплением рассматривает управление динамической системой как взаимодействие агента со средой: агент изменяет состояние среды, получает во время обучения за свои действия вознаграждение и стремится максимизировать суммарное вознаграждение за заданный период времени. В контексте космических полетов аппарат можно считать средой, программное обеспечение для управления – агентом, а вознаграждением может служить точность попадания в определенную область пространства и экономия затрат топлива. Стратегия управления (отображение состояния аппарата в управляющее воздействие) параметризуется, параметры оптимизируются таким образом, чтобы выполнялось уравнение оптимальности Беллмана, или чтобы вознаграждение, полученное агентом за весь полет, было максимальным в среднем по всем возможным начальным условиям. Результатом обучения является функция управления, которая способна привести космический аппарат в заданную область пространства. Эта функция может быть загружена на борт аппарата и применяться в режиме реального полета, используя текущее состояние аппарата или оценку этого состояния.

За несколько последних лет область обучения с подкреплением пополнилась эффективными алгоритмами, зарекомендовавшими себя в разных областях, в том числе и в механике космического полета (см. обзор литературы по теме в работе [17] в разделе Reinforcement learning). Эти численные методы основываются на алгоритмах приближенного динамического программирования, методах оптимизации функций с большим числом параметров и теории частично наблюдаемых марковских процессов принятия решений. Преимуществом этих методов является существенное сокращение математических предположений и значительный охват возможных решаемых задач. Примеры их применения показывают, что стратегии управления, создаваемые этими методами, естественным образом способны отображать навигационные измерения непосредственно в управляющие воздействия минуя фазу навигации и оценки состояния, а также адаптироваться к неизвестным параметрам аппарата и внешней среды, выходу из строя двигателей и ошибке в реактивной тяге [18–21]. Тем самым методы обучения с подкреплением выступают перспективным инструментом проектирования адаптивных автономных систем управления, навигации и управления.

Ранее была сформулирована методика приближенного решения задачи оптимального управления механической системой методами обучения с подкреплением [22]. Возникает интерес к применению указанной методики к задаче управления аппаратом во внешней области Солнечной системы в условиях неопределенности состояния аппарата и возможности выхода из строя управляющих органов движения.

Цель настоящей работы – разработка и исследование автономного управления для космического аппарата, движущегося в области фокуса гравитационной линзы Солнца, для получения изображений экзопланет или проведения других исследований с использованием методов обучения с подкреплением. Изучается возможность

использования видимых в телескоп изображений удаленных источников света, таких как экзопланеты, звезды, галактики, туманности, для расчета управляющих воздействий. Методы обучения с подкреплением используются для формирования законов управления аппаратом по оценкам состояния или непосредственно по наблюдениям с учетом неопределенности в движении аппарата и возможности выхода из строя двигателей. В настоящей работе исследованию подлежат сравнительно простые модели законов управления. Оптимизация же архитектуры моделей должна производиться с учетом специфических требований к миссии, этот вопрос в данной работе не рассматривается, однако полученные результаты исследования могут предоставить опорные значения характеристик движения для более совершенных моделей.

2. ПОСТАНОВКА ЗАДАЧИ АВТОНОМНОГО УПРАВЛЕНИЯ АППАРАТОМ В ОКРЕСТНОСТИ ФОКАЛЬНОЙ ЛИНИИ УДАЛЕННОГО ИСТОЧНИКА СВЕТА

Сформулируем задачу автономного управления движением центра масс космического аппарата в фокусе гравитационной линзы Солнца удаленного источника. Целью управления является перемещение аппарата на фокальную линию источника.

Сначала поясним некоторые термины гравитационного линзирования, которые будут использоваться в дальнейшем. Рассмотрим точечный источник света, расположенный на некотором достаточно большом расстоянии от Солнечной системы, чтобы пучок света можно было считать параллельным. Линия, проходящая через источник и Солнце, называется *фокальной линией солнечной гравитационной линзы* или просто *фокальной линией*, соответствующей источнику. Под действием гравитационного притяжения Солнца лучи света от источника изгибаются и фокусируются на фокальной линии (рис. 1). Точка на фокальной линии, в которую приходит луч света, зависит от расстояния, на котором луч проходит мимо Солнца: чем больше это расстояние, тем дальше от Солнца находится точка на фокальной линии. Так как световые лучи проходят на различных расстояниях от Солнца, но не ближе солнечного радиуса, весь свет от точечного источника фокусируется на геометрическом луче, начинающемся на расстоянии примерно 550 а.е. от Солнца [5]. Протяженные источники света (экзопланеты, звезды, галактики, туманности) можно рассматривать как состоящие из множества точечных источников. Каждой точке протяженного источника соответствует своя фокальная линия, поэтому весь свет от протяженного источника фокусируется не на геометрическом луче, а в некоторой области пространства, называемой *фокусом солнечной гравитационной линзы*, соответствующим протяженному источнику.

Допустим, в рамках некоторой миссии возникла задача автономной навигации аппарата по видимым в телескоп изображениям удаленных источников и задача нацеливания на фокальную линию солнечной гравитационной линзы. Пусть исследователями выбран удаленный источник света, по которому будет осуществляться навигация и к фокусу солнечной гравитационной линзы которого будет нацелено управление. В настоящей работе будем предполагать, что фокус, соответствующий источнику, имеет вид прямого кругового конуса, а под фокальной линией будем понимать ось этого конуса.

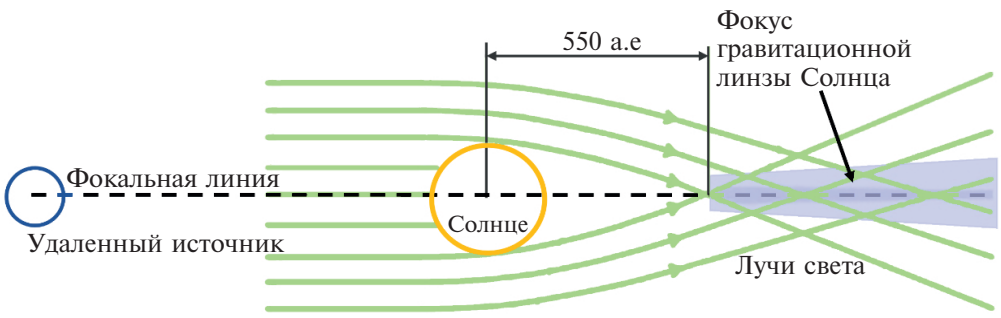

Рис. 1. Схематическое представление искривления лучей света удаленного источника под действием солнечной гравитационной линзы

Пусть космический аппарат в начальный момент времени находится на расстоянии R_0 от фокальной линии источника и движется со скоростью V_0 параллельно фокальной линии относительно некоторой квазиинерциальной системы отсчета, связанной с Солнцем. Предполагая параллельность скорости фокальной линии, мы выносим за пределы настоящей работы вопрос о предварительном выравнивании скорости и тем самым сокращаем число варьируемых параметров.

Ставится задача определения n_{imp} импульсов скорости, применяемых через одинаковые интервалы времени Δt_{imp}, которые через промежуток времени $(n_{\text{imp}}-1)\Delta t_{\text{imp}}$ приводят аппарат на фокальную линию. Каждый импульс скорости ограничен величиной $\Delta v_{\max}$. Считается, что расчет импульсов скорости производится на основании состояния аппарата, навигационной оценки этого состояния или по наблюдениям изображения в телескопе.

Будем считать, что движение аппарата между импульсами представляет собой равномерное и прямолинейное движение. Другими словами, внешние силы, действующие на аппарат, считаются пренебрежимо малыми. Это предположение можно считать разумным по следующим причинам. Во-первых, так как расстояние до Солнца превышает 550 а.е., ускорение силы притяжения к Солнцу не превышает $1.96 \cdot 10^{-8}$ м/с^2 = $= 0.146$ км/сут^2. Это значит, например, что за 30 дней ускорение притяжения Солнца сместит аппарат не более чем на 65.85 км. Во-вторых, это возмущение окажет влияние преимущественно на движение вдоль фокальной линии. Для наблюдения же удаленного источника определяющим является отклонение от фокальной линии, а не вдоль линии.

Сформулируем теперь поставленную выше задачу как задачу оптимального управления. Динамика механической системы описывается уравнениями

$$\dot{\mathbf{r}} = \mathbf{v}, \dot{\mathbf{v}} = 0,$$

где $\mathbf{r} \in \mathbb{R}^3$ – радиус-вектор аппарата, а $\mathbf{v} \in \mathbb{R}^3$ – вектор скорости аппарата в системе координат $Oxyz$, в которой начало O движется равномерно и прямолинейно вдоль фокальной линии со скоростью V_0, ось Ox направлена вдоль фокальной линии от Солнца, а оси Oy и Oz выбраны произвольным образом; $\mathbf{x} = [\mathbf{r}, \mathbf{v}] \in \mathbb{R}^6$ образует фазовый вектор аппарата. Движение системы рассматривается на фиксированном интервале времени $t \in [0, T]$. В моменты $t_i = i\Delta t_{\text{imp}}$, $i = 0, \ldots, n_{\text{imp}} - 1$, к системе прикладываются импульсы скорости, выражаемые функцией управления $\mathbf{u} = \mathbf{u}(\mathbf{x}) \in \mathbb{R}^3$ или $\mathbf{u} = \mathbf{u}(\mathbf{o}, \mathbf{h}) \in \mathbb{R}^3$, в зависимости от того, на основе какой информации производится управление – фазового вектора состояния системы $\mathbf{x}$ или наблюдения $\mathbf{o}$ и истории наблюдений $\mathbf{h}$, где под наблюдением и историей наблюдений могут пониматься оценки фазового вектора или непосредственно изображения. Финальный момент времени $T = n_{\text{imp}}\Delta t_{\text{imp}}$. Ограничения на фазовый вектор в процессе перелета не накладываются. Устанавливаются ограничения на функцию управления: $|\mathbf{u}| \leq \Delta v_{\max}$. Множество начальных условий Ω_0 определяется следующим образом:

$$\Omega_0 = \{\mathbf{x}_0 = [\mathbf{r}_0, \mathbf{v}_0] : \mathbf{r}_0 = [0, y_0, z_0],$$
$$\mathbf{v}_0 = [0,0,0], y_0^2 + z_0^2 \leq R_0^2\}. \quad (1)$$

Целевым множеством является

$$\Omega_{\text{targ}} = \{\mathbf{x} = [\mathbf{r}, \mathbf{v}] : \mathbf{r} = [x, 0, 0], \mathbf{v} = [v_x, 0, 0], x, v_x \in \mathbb{R}\}.$$

Целевым функционалом служит расстояние в фазовом пространстве до целевого множества в конце маневрирования:

$$\mathcal{J}(\mathbf{x}(\cdot), \mathbf{u}(\cdot)) = \sqrt{y_T^2 + z_T^2} + \alpha\sqrt{v_{yT}^2 + v_{zT}^2},$$

где y_T, z_T – это y- и z-координаты в момент $t = T$, а v_{yT} и v_{zT} – это y- и z-компоненты скорости в момент $t = T$; α – заданный весовой коэффициент нужной размерности.

Итак, ставится задача поиска функции управления $\mathbf{u} = \mathbf{u}(\mathbf{x})$ или $\mathbf{u} = \mathbf{u}(\mathbf{o}, \mathbf{h})$, которая для каждого начального условия $\mathbf{x}_0 \in \Omega_0$ порождает траекторию $(\mathbf{x}(t), \mathbf{u}(t))$, удовлетворяющую ограничению $|\mathbf{u}| \leq \Delta v_{\max}$ и минимизирующую функционал $\mathcal{J}$.

3. ПОСТАНОВКА ЗАДАЧИ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

Общая методика приближенного решения задачи оптимального управления методами обучения с подкреплением была ранее сформулирована и подробно рассмотрена в работе [22] на основе обзора литературы по приложениям методов машинного обучения в механике космического полета [17].

Для того чтобы решить задачу оптимального управления методом обучения с подкреплением, необходимо определить понятия теории обучения с подкреплением (состояние среды, действие агента, дискретный шаг среды, функции вознаграждения) в контексте поставленной задачи оптимального управления. В работе [22] эти понятия были рассмотрены для общей задачи оптимального управления, и процесс

формулирования задачи обучения с подкреплением был описан в шести шагах. Рассмотрим эти шаги в контексте исследуемой в настоящей работе задачи управления.

1. Под состоянием среды будем понимать фазовый вектор состояния аппарата $\mathbf{s} = \mathbf{x}$. В силу автономности системы уравнений движения фазовый вектор однозначно определяет эволюцию системы вне зависимости от значения момента времени, в который этот вектор задан.

2. На области Ω_0 начальных условий $\mathbf{x}_0$ определим равномерное распределение вероятностей $\mathcal{D}_0$, так как не предполагаем никакого приоритета одних начальных условий в области Ω_0 перед другими. Из этого распределения будут генерироваться начальные условия в серии эпизодов Монте-Карло для моделирования работы функции управления и эмпирической оценки функционала. Отметим, что после обучения стратегии на этапе ее тестирования или оценки качества, распределение начальных условий можно изменить и сделать их приближенными к ожидаемым в реальной миссии. В этом случае, впрочем, следует помнить, что распределение и математическое ожидание суммарных вознаграждений изменятся и оптимальная стратегия для нового распределения может отличаться от оптимальной стратегии для исходного распределения.

3. Дискретный шаг среды определим следующим образом:

$$\mathbf{x}'_k = \mathbf{x}_k + [0,0,0,\mathbf{u}_k], \mathbf{x}_{k+1} = \mathbf{x}'_k + \Delta t_{\text{imp}}\mathbf{f}(\mathbf{x}'_k),$$

где первое равенство означает применение импульса $\mathbf{u}_k$ на шаге k, $k = 0,\ldots,n_{\text{imp}} - 1$, а второе уравнение описывает изменение положения и скорости аппарата под действием динамики; здесь $\mathbf{f}$ – функция правой части уравнений движения. Второе равенство по сути является шагом метода Эйлера, но оно точно отражает динамику, так как уравнения движения являются линейными. Шаг $k = n_{\text{imp}} - 1$ и состояние $\mathbf{x}_{n_{\text{imp}}}$ считаются финальными. Последовательность из n_{imp} шагов назовем *эпизодом*.

4. Функцию вознаграждения определим следующим образом:

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

Суммарное вознаграждение за эпизод равно

$$R = \sum_{k=0}^{n_{\text{imp}}-1} r_k = \rho(\mathbf{x}_0) - \rho(\mathbf{x}_{n_{\text{imp}}}),$$

то есть улучшению расстояния до целевого множества за эпизод. Среднее значение именно этой величины будет максимизироваться в алгоритме обучения с подкреплением.

5. В качестве модели восприятия в настоящей работе рассматриваются три варианта. В первом варианте наблюдением является фазовый вектор аппарата, то есть $\mathbf{o} = \mathbf{x}$. Таким образом, изучается оптимизация и поведение стратегии управления аппаратом в условиях полной наблюдаемости и абсолютного знания состояния. Во втором варианте наблюдением выступает фазовый вектор с шумом, то есть $\mathbf{o} = \mathbf{x} + \xi$, где $\xi \in \mathcal{N}(0,\Sigma)$ – ошибка определения состояния аппарата, которая моделируется как нормально распределенный вектор с нулевым средним и задаваемой ковариационной матрицей Σ. Таким образом, в этом варианте моделируется *результат* процедуры навигации. В третьем варианте наблюдением $\mathbf{o}$ является модель изображения, поступающего на телескоп. Для моделирования видимого в телескоп изображения, получаемого в результате искривления лучей света под действием гравитации Солнца, существует множество программных инструментов, обзор которых можно найти в статье [23]. В настоящей работе используется адаптированная версия программы glafic2 (https://github.com/oguri/glafic2) [24], так как, в отличие от других программ, она находится в открытом доступе, а также создана и поддерживается специалистом в области гравитационного линзирования. Модель изображения представляет собой матрицу, каждая компонента которой принимает значения в интервале [0,1] и задается в оттенках серого. На рис. 2 и 3 продемонстрированы примеры

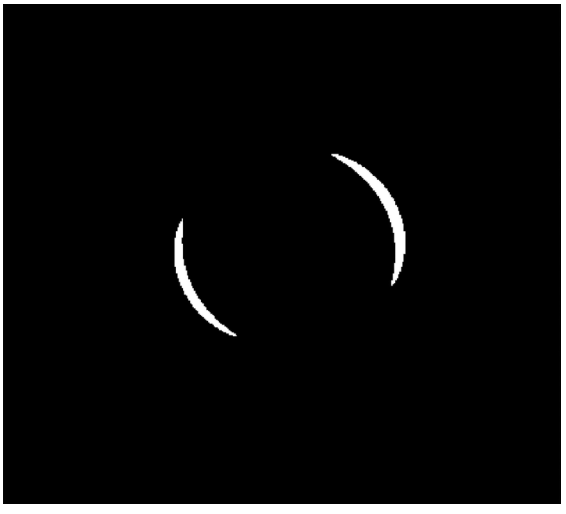

Рис. 2. Изображение, формируемое glafic2 для аппарата на расстоянии 130 тыс. км от фокальной линии и на расстоянии 600 а.е. от Солнца

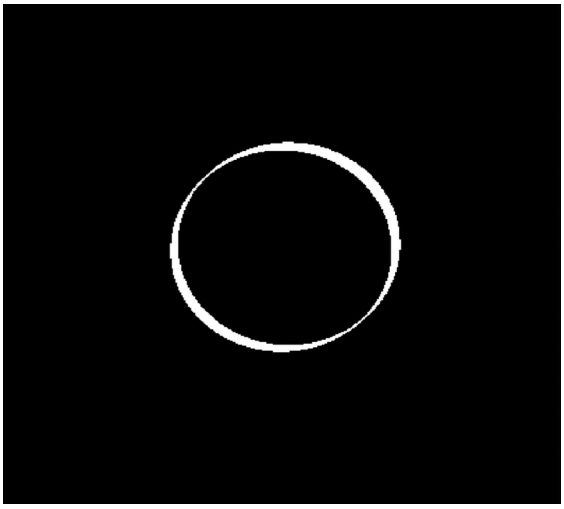

Рис. 3. Изображение, формируемое glafic2 для аппарата на расстоянии 85 тыс. км от фокальной линии и на расстоянии 600 а.е. от Солнца

изображений, формируемых glafic2 на расстояниях 130 и 85 тыс. км до фокальной линии соответственно (расстояние до Солнца – 600 а.е.).

6. Остается определить модель управления. В данной работе модель управления определяется с использованием классических для машинного обучения нейросетевых архитектур. Для $\mathbf{o} = \mathbf{x}$ и $\mathbf{o} = \mathbf{x} + \xi$ в качестве модели управления рассматривается нейросетевая модель с одним скрытым слоем размера $n_1 \geq 1$:

$$\mathbf{a} = \mathbf{A}_2^\pi \,\mathrm{th}(\mathbf{A}_1^\pi \mathbf{o} + \mathbf{b}_1^\pi) + \mathbf{b}_2^\pi, \mathbf{u} = \mathbf{a} \cdot \min(1, \Delta v_{\max} / |\,\mathbf{a}\,|), \quad (2)$$

где $\mathbf{a}$ – действие агента; $\mathbf{u}$ – управляющее воздействие; $\mathbf{o} \in \mathbb{R}^6$ – наблюдение; $\mathbf{A}_1^\pi \in \mathbb{R}^{n_1 \times 6}$, $\mathbf{b}_1^\pi \in \mathbb{R}^{n_1}$, $\mathbf{A}_2^\pi \in \mathbb{R}^{3 \times n_1}$, $\mathbf{b}_2^\pi \in \mathbb{R}^3$ – матрицы и векторы оптимизируемых параметров. Функция гиперболического тангенса действует на векторы покомпонентно[1]. Верхний индекс π означает, что параметры относятся к модели стратегии. В результате получается управляющее воздействие $\mathbf{u}$, по модулю не превосходящее $\Delta v_{\max}$.

Наблюдение-изображение связано с координатами аппарата, то есть его положением относительно фокальной линии, но не несет никакой информации о скорости движения аппарата. Два последовательных наблюдения-изображения, взятые в известные моменты времени t_1 и t_2, несут информацию не только об y- и z-координатах в эти моменты времени (y_1, x_1) и (y_2, x_2), но и о скорости движения $(v_{y1}, v_{z1}) = (v_{y2}, v_{z2}) = (y_2 - y_1, z_2 - z_1) / (t_2 - t_1)$, так как движение является равномерно прямолинейным. Поэтому два изображения несут информацию, достаточную для расчета импульса скорости для нацеливания на фокальную линию. В случае наблюдений-изображений рассматриваются две нейросетевые модели управления.

Во-первых, рассматривается модель со стеком входных данных

$$\mathbf{a} = \mathbf{A}_2^\pi \,\mathrm{th}(\mathbf{A}_1^\pi [\mathbf{o}, \mathbf{o}'] + \mathbf{b}_1^\pi) + \mathbf{b}_2^\pi,$$
$$\mathbf{u} = \mathbf{a} \cdot \min(1, \Delta v_{\max} / |\,\mathbf{a}\,|), \quad (3)$$

где $\mathbf{o} \in \mathbb{R}^{n_h n_w}$ и $\mathbf{o}' \in \mathbb{R}^{n_h n_w}$ – векторы изображений размера $n_h \times n_w$ с текущего и предыдущего шага; $[\mathbf{o}, \mathbf{o}'] \in \mathbb{R}^{2 n_h n_w}$ – конкатенация векторов $\mathbf{o}$ и $\mathbf{o}'$, $\mathbf{A}_1^\pi \in \mathbb{R}^{n_1 \times 2 n_h n_w}$, $\mathbf{b}_1^\pi \in \mathbb{R}^{n_1}$, $\mathbf{A}_2^\pi \in \mathbb{R}^{3 \times n_1}$, $\mathbf{b}_2^\pi \in \mathbb{R}^3$. В начальный момент времени $\mathbf{o}'$ не определен, для определенности этот вектор полагается нулевым. Наблюдение в предыдущий момент времени $\mathbf{o}'$ играет роль истории наблюдений. Такая нейросетевая модель выбрана исходя из следующих соображений. Все элементы матрицы изображения лежат в пределах от 0 (черный цвет) до 1 (белый цвет). Расположение единиц в матрице зависит от положения аппарата относительно фокальной линии, а количество единиц в матрице связано с близостью аппарата к фокальной линии. Следовательно, ожидается, что направление и величина импульса связаны линейными соотношениями с элементами матрицы и, возможно, одним нелинейным преобразованием. Поэтому в качестве нейросетевой архитектуры предлагается рассмотреть полносвязную сеть прямого распространения, а не сеть со сверточными слоями, применяемую обычно для распознавания образов.

Во-вторых, изучается рекуррентная сеть с простым рекуррентным слоем

$$\mathbf{h}_k = \mathrm{th}(\mathbf{A}_h^\pi \mathbf{o}_k + \mathbf{B}_h^\pi \mathbf{h}_{k-1} + \mathbf{c}_h^\pi), \mathbf{a}_k =$$
$$= \mathbf{A}_{a,2} \,\mathrm{th}(\mathbf{A}_{a,1}^\pi \mathbf{h}_{k-1} + \mathbf{b}_{a,1}^\pi) + \mathbf{b}_{a,2}^\pi, \quad (4)$$
$$\mathbf{u}_k = \mathbf{a}_k \cdot \min(1, \Delta v_{\max} / |\,\mathbf{a}_k\,|), k = 1, \ldots, n_{\mathrm{imp}} - 1,$$

где $\mathbf{h}_k \in \mathbb{R}^{n_1}$ – скрытое состояние рекуррентной сети, являющееся функцией текущего наблюдения и предыдущих наблюдений; n_1 – свободно выбираемый исследователем размер скрытого состояния; матрицы и векторы $\mathbf{A}_h^\pi \in \mathbb{R}^{n_1 \times n_h n_w}$, $\mathbf{B}_h^\pi \in \mathbb{R}^{n_1 \times n_1}$, $\mathbf{b}_h^\pi \in \mathbb{R}^{n_1}$, $\mathbf{A}_{a,1}^\pi \in \mathbb{R}^{n_2 \times n_1}$, $\mathbf{b}_{a,1}^\pi \in \mathbb{R}^{n_2}$, $\mathbf{A}_{a,2}^\pi \in \mathbb{R}^{3 \times n_2}$, $\mathbf{b}_{a,2}^\pi \in \mathbb{R}^3$ состоят из оптимизируемых параметров; n_2 – свободно выбираемый исследователем размер скрытого слоя. Начальное значение скрытого

[1] Здесь и далее функция гиперболического тангенса выбрана как одна из распространенных активационных функций в литературе по управлению космическими аппаратами с использованием методов обучения с подкреплением. Возможен выбор других активационных функций, например ReLU.

состояния $\mathbf{h}_0$ обычно полагается равным нулевому вектору.

Рекуррентные нейронные сети изначально создавались для обработки числовых последовательностей и зарекомендовали себя в задачах обработки временных рядов, естественного языка, распознавания речи. Они обладают способностью сохранять память о предыдущих входных данных, что позволяет им учитывать исторический контекст при составлении прогнозов. Рекуррентные сети могут обрабатывать последовательности произвольной длины. В то же время их обучение может быть ресурсоемким по сравнению с нейронными сетями прямого распространения, особенно при работе с длинными последовательностями, из-за проблемы исчезновения или увеличения градиента выхода сети по ее параметрам. Кроме того, рекуррентные сети могут испытывать проблемы с регистрацией долгосрочных зависимостей в данных. Впрочем, в контексте рассматриваемой задачи, когда для расчета величины и направления импульса необходимо и достаточно двух наблюдений-изображений, а эпизод состоит из небольшого числа шагов, эта проблема не играет большой роли.

Нейронные сети прямого распространения со стеком во входных данных обычно проще обучать и реализовывать, они не обладают проблемами исчезновения или увеличения градиентов, а данные при их использовании можно обрабатывать параллельно. Недостатком их использования становиться необходимость определения недостающих векторов наблюдения на первых итерациях. Задание вектора наблюдений нулевым вектором может приводить к ошибкам прогнозирования нейронной сети, учитывая, что нулевое значение наблюдения несет в себе содержательную информацию (в нашем случае – отсутствие наблюдения кольца Эйнштейна). При использовании нейронных сетей со стеками входных данных исследователи ожидают коррекции поведения системы после второй итерации.

Перейдем к вопросу оптимизации стратегии. В настоящее время наиболее распространенными методами оптимизации стратегии в случае непрерывных множеств состояния и действия являются метод градиента глубокой детерминированной стратегии (Deep Deterministic Policy Gradient, DDPG) [25], метод асинхронного исполнителя–критика (Asynchronous Advantage Actor Critic, A3C) [26], метод оптимизации ближайшей стратегии (Proximal Policy Optimization, PPO) [27], а также эволюционные алгоритмы [28, 29]. В механике космического полета хорошо зарекомендовал себя алгоритм PPO. По своей сути он является аналогом методов доверительных областей и был разработан для борьбы с неустойчивостью процесса обучения, когда небольшое изменение весов нейронной сети оказывает сильное влияние на стратегию, что дестабилизирует процесс обучения.

В настоящей работе оптимизация параметров функции управления производится с использованием реализации алгоритма PPO в рамках известной библиотеки stable-baselines3 (https://stable-baselines3.readthedocs.io/en/master/). Библиотека stable-baselines3 – одна из немногих профессиональных и открытых программных библиотек с алгоритмами обучения с подкреплением, реализованных специалистами в этой области. Разработка и сопровождение библиотеки ведется Немецким центром авиации и космонавтики (German Aerospace Center, DLR), Лабораторией интерактивной робототехники (Interactive Robotics Laboratory) в Университете Париж–Сакле, на Факультете электротехники и компьютерных наук (Electrical Engineering and Computer Science) в Калифорнийском университете и в Школе вычислительной техники (School of Computing) в Университете восточной Финляндии. Алгоритм PPO требует также задания модели для функции ценности. Во всех случаях формы наблюдений определим функцию ценности в рамках такой же архитектуры, как у модели управления – так часто делают в приложениях обучения с подкреплением. А именно, для случаев $\mathbf{o} = \mathbf{x}$ и $\mathbf{o} = \mathbf{x} + \xi$ возьмем

$$v = \mathbf{a}_2^{vT}\,\mathrm{th}(\mathbf{A}_1^v\mathbf{o} + \mathbf{b}_1^v) + b_2^v, \qquad (5)$$

где параметры $\mathbf{A}_1^v \in \mathbb{R}^{n_1\times 6}, \mathbf{b}_1^v \in \mathbb{R}^{n_1}, \mathbf{a}_2^v \in \mathbb{R}^{n_1}, b_2^v \in \mathbb{R}$, а верхний индекс v означает, что эти параметры относятся к модели функции ценности. В случае наблюдения-изображения, когда модель управления строится на основе стека входных данных, функцию ценности будем аппроксимировать функцией

$$v = \mathbf{a}_2^{vT}\,\mathrm{th}(\mathbf{A}_1^v[\mathbf{o},\mathbf{o}'] + \mathbf{b}_1^v) + b_2^v,$$

где $\mathbf{A}_1^v \in \mathbb{R}^{n_1\times 2n_hn_w}, \mathbf{b}_1^v \in \mathbb{R}^{n_1}, \mathbf{a}_2^v \in \mathbb{R}^{n_1}, b_2^v \in \mathbb{R}$. Наконец, если модель управления строится на основе рекуррентной сети, функцию ценности будем аппроксимировать как

$$\mathbf{h}_k = \mathrm{th}(\mathbf{A}_h^v\mathbf{o}_k + \mathbf{B}_h^v\mathbf{h}_{k-1} + \mathbf{b}_h^v),$$
$$v_k = \mathbf{a}_v^T\,\mathrm{th}(\mathbf{A}_v\mathbf{h}_{k-1} + \mathbf{b}_v) + b_v,$$

где $\mathbf{A}_h^v \in \mathbb{R}^{n_1\times n_hn_w}, \mathbf{B}_h^v \in \mathbb{R}^{n_1\times n_1}, \mathbf{b}_h^v \in \mathbb{R}^{n_1}, \mathbf{A}_v \in \mathbb{R}^{n_2\times n_1}$, $\mathbf{b}_v \in \mathbb{R}^{n_2}$, $\mathbf{a}_v^T \in \mathbb{R}^{n_2}$, $b_v \in \mathbb{R}$. Параметры n_1 и n_2

можно выбрать совпадающими с одноименными параметрами в модели управления.

На момент проведения исследования библиотека stable-baselines3 не позволяла создавать и обучать модели на основе рекуррентных сетей, используемых в настоящей работе. Ответвление от основного проекта SB3 Contrib (https://stable-baselines3.readthedocs.io/en/master/guide/sb3_contrib.html) имеет реализацию PPO на основе архитектуры длинной цепи элементов краткосрочной памяти (long short-term memory, LSTM) [30]. В настоящем же исследовании предпочтение отдается более простым нейросетевым моделям, поэтому мы адаптировали реализацию алгоритма PPO из SB3 Contrib для случая нашей рекуррентной сети.

Перейдем к описанию процедуры оптимизации при использовании алгоритма PPO. Задача состоит в оптимизации функционала

$$J(\theta) = \mathbb{E}\left(R \mid \pi(\theta)\right)$$

относительно вектора θ параметров стратегии $\pi(\theta)$. Теоретическое математическое ожидание в выражении для $J(\theta)$ заменяется на выборочное среднее. Для этого в соответствии с распределением начальных условий (см. пункт 2 выше) и дискретными шагами (пункт 3) производится серия испытаний Монте-Карло, агент на основании наблюдений (пункт 5) при фиксированных значениях параметров θ производит действия (пункт 6), получает за них вознаграждения r_k (пункт 4) и действует до конца эпизода. Так, в серии испытаний получаются реализации суммарных вознаграждений R за эпизод, их выборочное среднее дает оценку величины $J(\theta)$. Далее, на основании собранных данных (состояний, действий, вознаграждений) в серии испытаний алгоритм PPO корректирует параметры θ в сторону повышения значения функционала J, и процесс сбора данных повторяется снова. Процесс оптимизации останавливается, когда значение функционала перестает увеличиваться.

Поскольку оптимизация параметрической функции управления происходит на основе методов Монте-Карло и параметры моделей настраиваются исходя из опыта взаимодействия агента со средой, нет гарантии, что построенное управление будет решать поставленную задачу для любых начальных данных одинаково эффективно. Возникает задача оценки качества проектируемой системы управления. В настоящей работе предлагается оценивать качество с помощью неравенства Хефдинга [31, 22].

Теорема [31]. *Пусть $X_1, \ldots, X_n$ — независимые одинаково распределенные случайные величины, для которых выполнено $a \le X_i \le b$ с вероятностью единица. Тогда для среднего выборочного $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ справедлива оценка*

$$\mathbb{P}\left(\left|\bar{X} - \mathbb{E}X\right| \ge \varepsilon(b-a)\right) \le 2\exp\left(-2\varepsilon^2 n\right).$$

Следствие. *Если $X_i \in \{0,1\}$ имеют распределение Бернулли* $\mathrm{Be}(p_A)$*, то*

$$\mathbb{P}\left(\left|\bar{X} - p_A\right| \ge \varepsilon\right) \le 2\exp\left(-2\varepsilon^2 n\right).$$

Неравенство в утверждении теоремы называется *неравенством Хефдинга*, его удобно записывать в виде доверительного интервала:

$$\mathbb{E}X = \bar{X} \pm \varepsilon(b-a) \text{ с вероятностью не менее } 1 - 2\exp\left(-2\varepsilon^2 n\right).$$

Это неравенство означает, что, проведя n независимых измерений случайной величины X, ограниченной промежутком $[a, b]$, мы получим, что вероятность отклонения среднего выборочного $\bar{X}$ от истинного математического ожидания $\mathbb{E}X$ более чем на $\varepsilon(b-a)$ не превосходит $p = 2\exp(-2\varepsilon^2 n)$. Из неравенства Хефдинга следует, например, что, проведя $n = 3800452$ измерений случайной величины X, мы получим, что вероятность отклонения среднего выборочного $\bar{X}$ от истинного математического ожидания $\mathbb{E}X$ более чем на $10^{-3}(b-a)$ не превышает $p = 0.1\%$. Следствие из теоремы позволяет оценивать вероятность p_A наступления события A, которое, в зависимости от исхода, может происходить или не происходить. В настоящей работе нас будут интересовать оценки истинных средних промаха по положению мимо фокальной линии, промаха по скорости, характеристические затраты топлива, а также вероятность наступления события, которое заключается в том, что конечный промах по положению больше начального промаха.

Итак, поставлена задача обучения с подкреплением — описаны варианты состояния среды (космического аппарата) и их начальные значения, действия агента (управляющей программы), переход между состояниями, функция вознаграждения, модель восприятия, а также параметрические модели управления, алгоритм их оптимизации и способ оценки их качества. Перейдем теперь к решению поставленной задачи.

4. РЕШЕНИЕ ЗАДАЧИ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

В разделе приводятся результаты исследования затрат характеристической скорости для нацеливания на фокусную линию, оценки финальной точности нацеливания и качества работы функции управления.

Случай 1. Наблюдением является состояние аппарата

Начнем со случая, когда наблюдением является состояние аппарата, то есть $\mathbf{o} = \mathbf{x}$. Для определенности будем полагать, что область начальных условий (1) представляет собой круг радиуса $R_0 = 100$ тыс. км. Движение начинается на расстоянии 550 а.е. от Солнца. Эпизод длится 30 дней, число импульсов $n_{\text{imp}} = 6$, интервал между импульсами $\Delta t_{\text{imp}} = 5$ дней. Импульсы ограничим величиной $\Delta v_{\max} = 100$ м/с. Положение и скорость аппарата определяются в безразмерной системе единиц: единица расстояния равна 100 тыс. км, единица времени равна 1 км/с. Весовой коэффициент α в функции вознаграждения равен единице в безразмерной системе единиц.

Обучение моделей произведем с помощью реализации метода PPO из библиотеки stable-baselines3. Число нейронов на скрытом слое в модели управления (2) и модели функции ценности (5) выберем равным $n_1 = 6$. Объем выборки для аппроксимации среднего значения функционала J (опция n_steps) выберем равным 10000. Число итераций градиентного метода для коррекции весов нейросетевых моделей (опция n_epochs) выберем равным 30, а скорость обучения (опция learning_rate) – 0.001. Обучение будем проводить на центральном процессоре и завершим, когда число дискретных шагов достигнет 1.5 млн (опция total_timesteps). Здесь и в других случаях опции подбирались вручную до тех пор, пока не начинала наблюдаться “сходимость” параметров нейросетевых моделей, которая проявляла себя в том, что в среднем значение функционала сначала монотонно росло, а затем переставало изменяться. Сходимость наблюдалась в широком диапазоне значений опций. Их оптимизация в настоящей работе не рассматривалась.

Результаты обучения модели управления и функции ценности оказались следующими. Оценка среднего суммарного вознаграждения выросла со значения –0.3066 на первых итерациях до 0.6505 на последней итерации. Заметим, что для теоретически оптимальной стратегии, которая полностью устраняет невязку по положению и скорости относительно фокальной линии и тем самым максимизирует математическое ожидание суммарного вознаграждения, среднее значение суммарного вознаграждения равно

$$\int_0^{R_0}\int_0^{2\pi}\frac{r^2}{\pi}drd\varphi = \int_0^1 2r^2\,dr = \frac{2}{3} \approx 0.6667,$$

что говорит о близости найденного управления к оптимальному. Важно отметить, что точный расчет математического ожидания оказался возможным благодаря удачному выбору функции вознаграждения, виду суммарного вознаграждения как телескопической суммы вознаграждений за эпизод, и тому, что оптимальное среднее суммарное вознаграждение не требует в нашем случае определения теоретически оптимального управления.

Наличие промаха по положению и скорости в конце эпизода объясняется ограниченными возможностями нейросетевой архитектуры и сходимостью оптимизационной процедуры лишь к локальному минимуму функционала, который, в свою очередь, тоже оценивается приближенно в серии испытаний Монте-Карло.

Качество полученного управления было протестировано в 3800452 испытаниях Монте-Карло при равномерном распределении на Ω_0 начальных условий. Результаты оценок промаха по положению и скорости в конце эпизода, а также суммарные затраты характеристической скорости приведены в табл. 1. Здесь q_0 означает минимальное значение величины из встреченных; $q_{0.25}$ – величину, ниже которой находятся 25 % встреченных величин; $q_{0.5}$ – медиану; $q_{0.75}$ означает величину, выше которой находятся 25 % встреченных величин; q_1 – максимальное значение величины из встреченных; μ – среднее арифметическое всех величин. Из таблицы следует, что в среднем промах мимо фокальной линии составляет около 1215 км, причем худшее значение промаха равно примерно 2650 км и намного меньше максимального начального промаха в 100 тыс. км. Под промахом по скорости понимается скорость в плоскости yz, в среднем она равна $\sim$ 17 м/с, а в худшем случае – 30 м/с. Средние затраты характеристической скорости равны $\sim$ 51 м/с.

На рис. 4 показаны начальные и конечные промахи по положению мимо фокальной линии. Из рисунка видно, что почти во всех исходах расстояние до линии уменьшается (почти все точки находятся под красной линией, отмечающей равные начальные и конечные расстояния). Увеличение

Таблица 1. Квантили и средние значения распределений промаха по положению Δr_f и скорости Δv_f мимо фокальной линии и суммарные за эпизод затраты характеристической скорости u для стратегии, основанной на состоянии аппарата

Показатель	q_0	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{1.0}$	μ
Δr_f, км	1.2028	651.8684	1096.0418	1723.1587	2650.0928	1214.8167
Δv_f, м/с	0.0125	11.9971	17.4724	21.9254	29.6690	16.8165
u, м/с	0.2695	39.0842	54.1995	65.1141	80.3756	50.9032

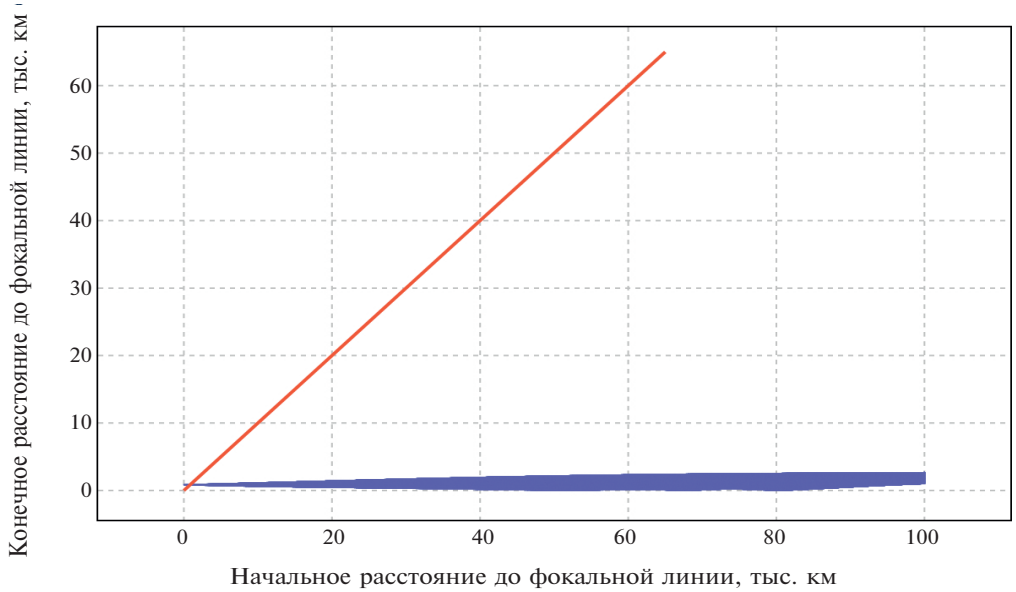

Рис. 4. Начальные и конечные расстояния (точки синего цвета) до фокальной линии, полученные в серии испытаний Монте-Карло, для стратегии, основанной на состоянии аппарата; красным цветом обозначена линия равенства расстояний

расстояния наблюдалось только для исходов с начальным расстоянием до линии до 900 км.

Из неравенства Хефдинга и расчетов следует, что истинное среднее значение вероятности неблагоприятного исхода (увеличение невязки по положению по сравнению с начальным значением невязки) равно 0.007 % ± 0.1 % с вероятностью не менее 99.9 % и превышает 0.017 % с вероятностью не более 0.05 %. При условии $\Delta r_f < 100$ тыс. км истинное среднее значение промаха по положению лежит в интервале 1215 ± 100 км с вероятностью не менее 99.9 % Учитывая, что априори $\Delta v_f < 600$ м/с, получаем, что истинное среднее значение промаха по скорости лежит в интервале 16.8 ± 0.6 м/с с вероятностью не менее 99.9 %. Наконец, так как суммарные затраты характеристической скорости априори не превышают $6\Delta v_{\max} = 600$ м/с, истинное среднее значение затрат характеристической скорости лежит в интервале 50.9 ± 0.6 м/с с вероятностью не менее 99.9 %

Случай 2. Наблюдением является оценка состояния аппарата

Рассмотрим случай $\mathbf{o} = \mathbf{x} + \xi$, где ξ – нормальный случайный вектор с нулевым математическим ожиданием и диагональной матрицей ковариаций Σ. Дисперсии по каждой координате, для примера, выберем равными $\sigma_r = 10$ тыс. км, а по каждой компоненте скорости – $\sigma_v = 0.3$ м/с. Параметры моделей и опции библиотеки обучения возьмем теми же самыми, что и в предыдущем подразделе.

Оценка среднего суммарного вознаграждения во время обучения выросла со значения 0.06224 на первых итерациях до 0.5999 на последней итерации. Это значение меньше, чем в случае $\mathbf{o} = \mathbf{x}$ (0.6505), и является следствием влияния ошибки знания состояния.

Качество полученного управления было протестировано в 3800452 испытаниях Монте-Карло при равномерном распределении на Ω_0 начальных условий. Результаты оценок

промаха по положению и скорости в конце эпизода, а также суммарные затраты характеристической скорости приведены в табл. 2. Из таблицы следует, что в среднем промах мимо фокальной линии составляет около 5493 км, причем худшее значение промаха равно примерно 24716 км. Эти значения в несколько раз выше значений, полученных для случая $\mathbf{o} = \mathbf{x}$. Повышение промаха происходит вследствие ошибки по положению порядка 10–20 тыс. км, в результате чего агент не способен научиться принимать верное решение об управляющем воздействии. Средний промах по скорости вырос незначительно: 17.8 м/с. В худшем случае наблюдался промах по скорости величиной 45.4 м/с. Средние затраты характеристической скорости равны приблизительно 58.9 м/с.

На рис. 5 показаны начальные и конечные промахи по положению мимо фокальной линии. Из рисунка видно, что во всех исходах, когда начальное расстояние до линии фокуса превышало 20 тыс. км, промах по положению в конце эпизода уменьшался, но до порядка 20 тыс. км.

Из неравенства Хефдинга и расчетов следует, что с вероятностью не менее 99.9 % истинное среднее значение вероятности неблагоприятного исхода (увеличение невязки по положению по сравнению с начальным значением невязки) равно 0.3911 % $\pm$ 0.1 %, истинное среднее значение промаха по положению (при условии $\Delta r_f < 100$ тыс. км) равно 5493 $\pm$ 100 км, истинное среднее значение промаха по скорости равно 17.8 $\pm$ 0.6 м/с, а истинное среднее значение затрат характеристической скорости равно 58.9 $\pm$ 0.6 м/с.

Случай 3. Наблюдением является изображение

Рассмотрим теперь случай, когда наблюдением является изображение. В настоящей работе изображение моделируется с использованием программной библиотеки гравитационного линзирования glafic2. Размеры изображения – 20

Таблица 2. Квантили и средние значения распределений промаха по положению Δr_f и скорости Δv_f мимо фокальной линии и суммарные за эпизод затраты характеристической скорости u для стратегии, основанной на оценке состояния аппарата

Показатель	q_0	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{1.0}$	μ
Δr_f, км	1.9677	3319.3015	5155.4847	7297.5143	24715.8587	5492.7378
Δv_f, м/с	0.0094	12.4867	18.1932	23.2020	45.3685	17.7810
u, м/с	7.4632	48.6842	60.3028	69.6961	117.7929	58.8619

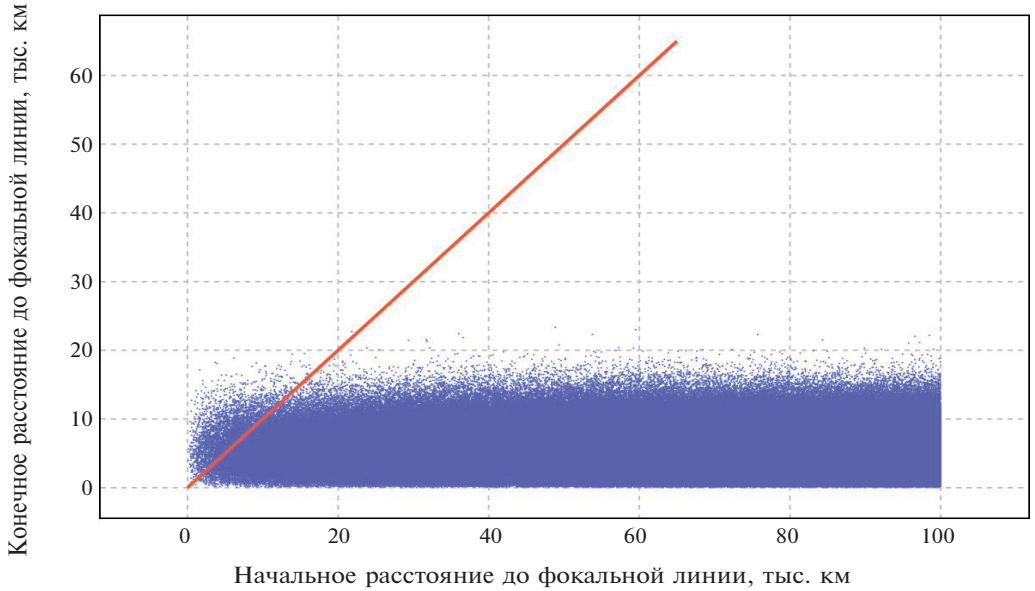

Рис. 5. Начальные и конечные расстояния (точки синего цвета) до фокальной линии, полученные в серии испытаний Монте-Карло, для стратегии, основанной на оценке состояния аппарата

пикселей в высоту и 20 в ширину. Цвета – оттенки серого, один канал.

Разберем случай, когда в качестве модели управления используется модель со стеком входных данных (3). Здесь $n_h = 20$, $n_w = 20$, а число нейронов на промежуточном слое выбрано равным $n_1 = 8$. Объем выборки для аппроксимации среднего значения функционала J (опция n_steps) выбран равным 5000 с размером пакета 1250 (опция batch_size). Число итераций градиентного метода для коррекции весов нейросетевых моделей (опция n_steps) выберем равным 10, а скорость обучения (опция learning_rate) – 0.001. Обучение будем производить на центральном процессоре и завершим, когда число дискретных шагов достигнет 500 тыс. (опция total_timesteps).

Оценка среднего суммарного вознаграждения во время обучения выросла со значения –0.1928 на первых итерациях до 0.5663 на последней итерации. Это значение меньше, чем в случае $\mathbf{o} = \mathbf{x}$ (0.6505) и в случае $\mathbf{o} = \mathbf{x} + \xi$ (0.5999).

Качество полученного управления было протестировано в 152019 испытаниях Монте-Карло при равномерном распределении на Ω_0 начальных условий. Результаты оценок промаха по положению и скорости в конце эпизода, а также суммарные затраты характеристической скорости приведены в табл. 3. Из таблицы следует, что в среднем промах мимо фокальной линии составляет около 7631 км, причем худшее значение промаха равно ~26222 км. Эти значения примерно на 2000 км больше соответствующих значений для случая $\mathbf{o} = \mathbf{x} + \xi$. Средний промах по скорости: 29.7 м/с, а в худшем случае наблюдался промах по скорости величиной 58.2 м/с. Эти значения примерно на 12 м/с больше соответствующих значений для случая $\mathbf{o} = \mathbf{x} + \xi$. Средние затраты характеристической скорости равны ~46.3 м/с, что также примерно на 12 м/с больше соответствующего значения для $\mathbf{o} = \mathbf{x} + \xi$.

На рис. 6 показаны начальные и конечные промахи по положению мимо фокальной линии. Четкие структуры при малых значениях начального

Таблица 3. Квантили и средние значения распределений промаха по положению Δr_f и скорости Δv_f мимо фокальной линии и суммарные за эпизод затраты характеристической скорости u для стратегии, основанной на стеке изображений

Показатель	q_0	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{1.0}$	μ
Δr_f, км	8.4155	4712.2326	7067.1193	10034.0078	26222.5645	7531.1550
Δv_f, м/с	0.8226	18.8295	30.1019	41.2399	58.2252	29.7046
u, м/с	16.0063	38.2686	46.2525	54.6776	75.8273	46.3166

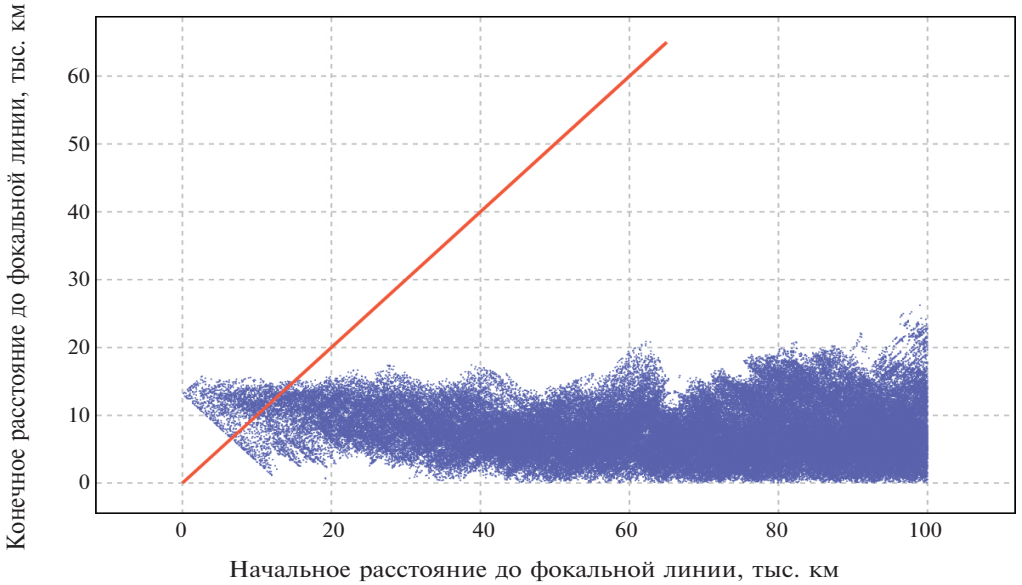

Рис. 6. Начальные и конечные расстояния (точки синего цвета) до фокальной линии, полученные в серии испытаний Монте-Карло, для стратегии, основанной на стеке изображений

промаха связаны с совпадающими наблюдениями: используемая в данном примере модель изображений колец Эйнштейна не позволяет различить изображения вблизи фокальной линии.

Из неравенства Хефдинга и расчетов следует, что с вероятностью не менее 99.9 % истинное среднее значение вероятности неблагоприятного исхода (увеличение невязки по положению по сравнению с начальным значением невязки) равно 1.16 % ± 0.5 %, истинное среднее значение промаха по положению (при условии $\Delta r_f < 100$ тыс. км) составляет 7531 ± 500 км, истинное среднее значение промаха по скорости равно 29.7 ± 3 м/с, а истинное среднее значение затрат характеристической скорости достигает 46.3 ± 3 м/с.

Перейдем теперь к случаю модели управления с рекуррентной сетью (4), вместо стека входных данных. Размер скрытого состояния n_1 также был выбран равным 8. Качество обученного управления было протестировано в 152019 испытаниях Монте-Карло при равномерном распределении на Ω_0 начальных условий. Результаты оценок промаха по положению и скорости в конце эпизода, а также суммарные затраты характеристической скорости приведены в табл. 4. Из таблицы следует, что в среднем промах мимо фокальной линии составляет около 6012 км, причем худшее значение промаха равно примерно 27882 км. Средний промах по скорости: 29.1 м/с, а в худшем случае наблюдался промах по скорости величиной 60.6 м/с. Средние затраты характеристической скорости равны приблизительно 50.7 м/с. Все эти значения близки к соответствующим значениям, полученным для модели управления со стеком входных данных. Доверительные интервалы для промахов и характеристической скорости также меняются незначительно, по сравнению с предыдущей моделью. На рис. 7 показаны начальные и конечные промахи по положению мимо фокальной линии. Картина также мало отличается от предыдущей, демонстрируемой на рис. 6.

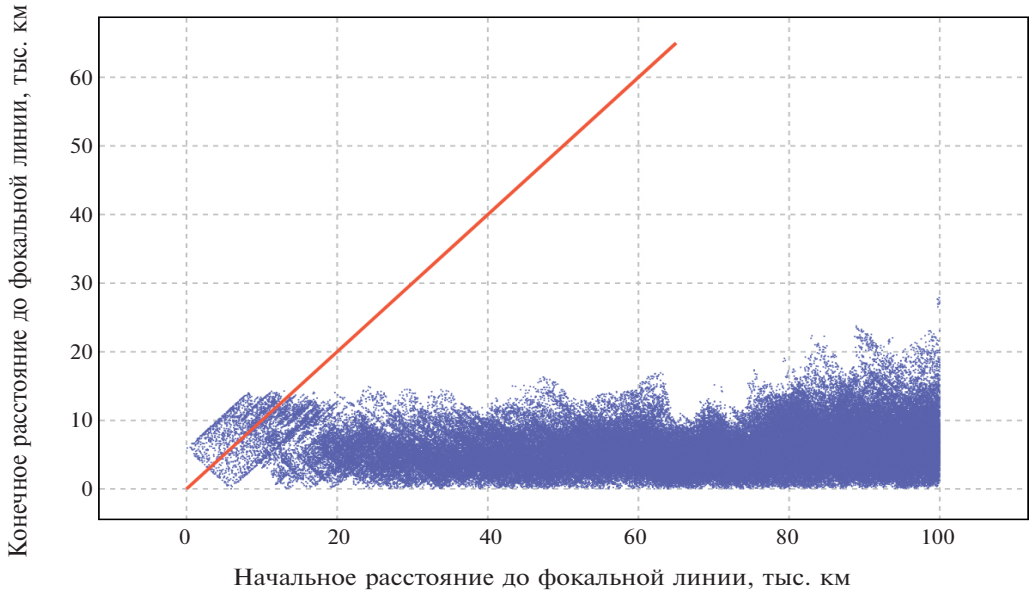

Рис. 7. Начальные и конечные расстояния (точки синего цвета) до фокальной линии, полученные в серии испытаний Монте-Карло, для стратегии, основанной на рекуррентной сети

Таблица 4. Квантили и средние значения распределений промаха по положению Δr_f и скорости Δv_f мимо фокальной линии и суммарные за эпизод затраты характеристической скорости u для стратегии, основанной на рекуррентной сети

Показатель	q_0	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{1.0}$	μ
Δr_f, км	14.1111	3613.3095	5555.9305	7897.0397	27881.7771	6011.7537
Δv_f, м/с	0.0664	17.1451	29.3133	40.7824	60.6289	29.0621
u, м/с	9.5444	43.6363	50.8189	58.5343	88.0539	50.6818

5. ОБУЧЕНИЕ С УЧЕТОМ ОШИБОК ИСПОЛНЕНИЯ МАНЕВРОВ

Перейдем к вопросу построения автономного управления, адаптивного к ошибкам исполнения импульсов скорости. Ошибку будем моделировать следующим образом. В начале каждого эпизода реализуется случайная величина $\rho \in [0.5,1.6]$ с непрерывным равномерным распределением. Эта величина не известна агенту. Эволюция состояний среды происходит таким образом, что вместо применения импульса $\mathbf{u}$ на основе действия агента применяется импульс $\rho\mathbf{u}$. Эта ошибка может быть связана, например, с уменьшенным или увеличенным интервалом действия тяги или с неверным уровнем силы тяги. Необходимо получить управление, способное адаптироваться к ошибкам исполнения маневров, основываясь на наблюдениях состояния аппарата.

Рассмотрим два случая наблюдений – с оценкой состояния $\mathbf{o} = \mathbf{x} + \xi$ и изображением. Параметры модели среды и управления и процедуры обучения выбраны теми же самыми, что и в предыдущем разделе. В результате обучения были получены агенты, обученные в ситуациях с ошибкой по величине импульса до 30 %.

В случае $\mathbf{o} = \mathbf{x} + \xi$ оценка среднего суммарного вознаграждения во время обучения выросла со значения 0.1405 на первых итерациях до 0.5725 на последней итерации, что немного меньше, чем в ситуации без ошибок исполнения импульсов (0.5999). В табл. 5 показаны результаты моделирования обученной стратегии для значений ρ от 0.5 до 1.6, указаны средние значения промаха по положению и по скорости, а также средние значения затрат характеристической скорости на основе действий агента; число испытаний Монте-Карло для каждого значения ρ равнялось 152019, что согласно неравенству Хефдинга позволяет получать доверительный интервал значений оцениваемых параметров размера $0.01(b - a)$ на уровне доверия 99.9 %, где a и b – априорные минимальное и максимальное значения оцениваемого параметра. Значения промаха по положению округлены до первого знака после запятой, а промах по скорости и характеристическая скорость – до целых. Из таблицы видно, что промах по положению остается одинаковым на интервале $\rho \in [0.8,1.2]$, а для значений, на которых агент не обучался, невязки по положению растут тем сильнее, чем дальше ρ находится от интервала. Рост невязки по скорости объясняется тем, что в функции вознаграждения часть, связанная с промахом по положению, оказывается больше части, связанной с промахом по скорости, в результате чего агент больше стремится понижать невязку по положению, но не по скорости.

В случае наблюдений-изображений оценка среднего суммарного вознаграждения во время обучения выросла со значения –0.1303 на первых итерациях до 0.5663 на последней итерации, что близко к случаю $\mathbf{o} = \mathbf{x} + \xi$. Аналогичные результаты моделирования обученной стратегии для значений ρ от 0.6 до 1.6 показаны в табл. 6. Здесь для каждого фиксированного значения ρ производилось 265 испытаний Монте-Карло, что, согласно неравенству Хефдинга, позволяет получать доверительный интервал значений оцениваемых параметров размера $0.2(b - a)$ на уровне доверия 99.9 %, где a и b – априорные

Таблица 5. Средние значения промаха по положению и скорости и средние значения характеристической скорости для различных значений ошибки исполнения импульсов для управления по оценке состояния

ρ	1.6	1.5	1.4	1.3	1.2	1.1	1.0	0.9	0.8	0.7	0.6	0.5
$\mu_{\Delta r_f}$, 10^3 км	6.2	6.0	5.8	5.7	5.6	5.6	5.6	5.6	5.7	6.3	8.1	11.4
$\mu_{\Delta v_f}$, м/с	11	11	11	12	13	14	16	18	20	22	24	25
μ_u, м/с	62	63	63	64	64	65	65	66	66	66	68	70

Таблица 6. Средние значения промаха по положению и скорости и средние значения характеристической скорости для различных значений ошибки исполнения импульсов для управления по изображениям

ρ	1.6	1.5	1.4	1.3	1.2	1.1	1.0	0.9	0.8	0.7	0.6
$\mu_{\Delta r_f}$, 10^3 км	17.2	15.8	13.7	11.6	9.7	7.4	6.4	7.3	11.4	15.0	21.0
$\mu_{\Delta v_f}$, м/с	20	21	22	21	22	22	24	27	26	23	22
μ_u, м/с	53	53	51	51	52	51	50	49	48	47	47

минимальное и максимальное значения оцениваемого параметра. Из табл. 6 видно, что чем дальше находится значение ρ от 1.0, тем больше средний промах по положению.

6. ЗАКЛЮЧЕНИЕ

В настоящей работе выведены и исследованы автономные квазиоптимальные законы управления космическим аппаратом в области фокуса гравитационной линзы Солнца, построенные с использованием методов обучения с подкреплением. Задача управления состояла в нацеливании на фокальную линию условной экзопланеты за несколько воздействий импульса скорости при старте на расстоянии до 100 тыс. км до фокальной линии с параллельной ей скоростью.

В результате расчетов были получены следующие выводы. При построении отображения положения и скорости аппарата в управляющие воздействия промах по положению варьируется в пределах от 1 до 2700 км, промах по скорости – от 0.012 до 30 м/с, а суммарные затраты характеристической скорости – от 0.27 до 81 м/с. В случае отображения в управляющие воздействия оценки состояния аппарата со среднеквадратическими ошибками по положению 10 тыс. км и по скорости 0.3 м/с, финальный промах по положению лежит в пределах от 2 км до 25 тыс. км, по скорости – от 0.01 до 46 м/с, а затраты характеристической скорости – от 7 до 118 м/с. Законы управления на основе моделей изображений колец Эйнштейна приводят к промаху по положению от 8 км до 28 тыс. км, по скорости – от 0.1 до 60 м/с, и к затратам характеристической скорости от 10 до 76 м/с. Архитектуры на основе стека входных данных и рекуррентной сети дают близкие результаты. В работе также показано, что модели управления могут быть успешно обучены и в ситуациях, когда двигатель аппарата производит тягу не в полной мере или избыточно.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа поддержана грантом Российского научного фонда (проект № 22-71-00051).

КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет, что у него нет конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Brandt P.C., Provornikova E.A., Cocoros A. et al.* Interstellar Probe: Humanity's exploration of the Galaxy Begins // Acta Astronautica. 2022. V. 199. P. 364–373. https://doi.org/10.1016/j.actaastro.2022.07.011
2. *Einstein A.* The Field Equations of Gravitation // Preussische Akademie der Wissenschaften, Sitzungsberichte, (Math. Phys.). Berlin, 1915. P. 844–847.
3. *Eddington A.S.* Space, time and gravitation: An outline of the general relativity theory. Cambridge University Press, 1920.
4. *Фок В.А.* Теория пространства, времени и тяготения. Москва: Физматгиз, 1955.
5. *Turyshev S.G., Toth V.T.* Resolved imaging of exoplanets with the solar gravitational lens // Monthly Notices of the Royal Astronomical Society. 2022. V. 515. Iss. 4. P. 6122–6132. https://doi.org/10.1093/mnras/stac2130
6. *Turyshev S.G.* Wave-theoretical description of the solar gravitational lens // Physical Review. 2017. V. 95. Iss. 8. Art. ID. 084041. https://doi.org/10.1103/PhysRevD.95.084041
7. *Turyshev S.G., Toth V.T.* Wave-optical treatment of the shadow cast by a large gravitating sphere // Physical Review. 2018. V. 98. Iss. 10. Art. ID. 104015. https://doi.org/10.1103/PhysRevD.98.104015
8. *Turyshev S.G., Toth V.T.* Optical properties of the solar gravitational lens in the presence of the solar corona // European Physical J. Plus. 2019. V. 134. Art. ID. 63. https://doi.org/10.1140/epjp/i2019-12426-4
9. *Turyshev S.G., Toth V.T.* Image formation for extended sources with the solar gravitational lens // Physical Review. 2020. V. 102. Iss. 2. Art. ID. 024038. https://doi.org/10.1103/PhysRevD.102.024038
10. *Toth V.T., Turyshev S.G.* Image recovery with the solar gravitational lens // Physical Review. 2021. V. 103. Iss. 12. Art. ID. 124038. https://doi.org/10.1103/PhysRevD.103.124038
11. *Willems P.A.* Photometric Limits on the High Resolution Imaging of Exoplanets Using the Solar Gravity Lens // Acta Astronautica. 2018. V. 152. P. 408–414. https://doi.org/10.1016/j.actaastro.2018.08.013
12. *Turyshev S.G., Shao M., Alkalai L. et al.* Direct Multipixel Imaging and Spectroscopy of an exoplanet with a Solar Gravity Lens Mission // Final Report. NASA Innovative Advanced Concepts (NIAC). Phase I. 2018. https://arxiv.org/abs/1802.08421
13. *Turyshev S.G., Shao M., Toth V.T. et al.* Direct Multipixel Imaging and Spectroscopy of an Exoplanet with a Solar Gravity Lens Mission // Final Report. NASA Innovative Advanced Concepts (NIAC). Phase II. 2020. https://arxiv.org/abs/2002.11871
14. *Саттон Р.С., Барто Э.Г.* Обучение с подкреплением. Москва: Бином. Лаборатория знаний, 2017.

15. *Bertsekas D.P.* Reinforcement learning and optimal control. Belmont: Athena Scientific, 2019.

16. *Kamalapurkar R., Walters P., Rosenfeld J. et al.* Reinforcement Learning for Optimal Feedback Control. A Lyapunov-Based Approach. Cham: Springer, 2018.

17. *Shirobokov M., Trofimov S., Ovchinnikov M.* Survey of machine learning techniques in spacecraft control design // Acta Astronautica. 2021. V. 186. P. 87–97. https://doi.org/10.1016/j.actaastro.2021.05.018

18. *Gaudet B., Linares R., Furfaro R.* Terminal adaptive guidance via reinforcement meta-learning: Applications to autonomous asteroid close-proximity operations // Acta Astronautica. 2020. V. 171. P. 1–13. https://doi.org/10.1016/j.actaastro.2020.02.036

19. *Gaudet B., Linares R., Furfaro R.* Adaptive guidance and integrated navigation with reinforcement meta-learning // Acta Astronautica. 2020. V. 169. P. 180–190. https://doi.org/10.1016/j.actaastro.2020.01.007

20. *Scorsoglio A., D'Ambrosio A., Ghilardi L. et al.* Image-based deep reinforcement meta-learning for autonomous lunar landing // J. Spacecraft and Rockets. 2022. V. 59. Iss. 1. P. 153–165. https://doi.org/10.2514/1.A35072

21. *Gaudet B., Linares R., Furfaro R.* Six degree-of-freedom body-fixed hovering over unmapped asteroids via LIDAR altimetry and reinforcement meta-learning // Acta Astronautica. 2020. V. 172. P. 90–99. https://doi.org/10.1016/j.actaastro.2020.03.026

22. *Широбоков М.Г.* Методика построения управления космическими аппаратами с использованием методов обучения с подкреплением // Косм. исслед. 2024. Т. 62. № 5. С. 498–515. https://doi.org/10.31857/S0023420624050082

23. *Lefor A.T., Futamase T., Akhlaghi M.* A systematic review of strong gravitational lens modeling software // New Astronomy Reviews. 2013. V. 57. Iss. 1–2. P. 1–13. https://doi.org/10.1016/j.newar.2013.05.001

24. *Oguri M.* The Mass Distribution of SDSS J1004+4112 Revisited // Public. Astronomical Society of Japan. 2010. V. 62. Iss. 4. P. 1017–1024. https://doi.org/10.1093/pasj/62.4.1017

25. *Silver D., Lever G., Heess N. et al.* Deterministic policy gradient algorithms // Proc. 31st Intern. Conf. Machine Learning. Beijing, China. 2014. V. 32. Iss. 1. P. 387–395. http://proceedings.mlr.press/v32/silver14.html

26. *Mnih V., Badia A.P., Mirza M. et al.* Asynchronous Methods for Deep Reinforcement Learning // Proc. 33rd Intern. Conf. Machine Learning. New York, USA. 2016. V. 48. P. 1928–1937. https://proceedings.mlr.press/v48/mniha16.html

27. *Schulman J., Wolski F., Dhariwal P. et al.* Proximal Policy Optimization Algorithms // arXiv preprint. 2017. Art. ID. 1707.06347. https://arxiv.org/abs/1707.06347

28. *Moriarty D.E., Schultz A. C., Grefenstette J.J.* Evolutionary algorithms for reinforcement learning // J. Artificial Intelligence Research. 1999. V. 11. P. 241–276.

29. *Sehgal A., La H., Louis S. et al.* Deep reinforcement learning using genetic algorithm for parameter optimization // Proc. Third IEEE International Conference on Robotic Computing (IRC). 2019. P. 596–601. https://doi.org/10.1109/IRC.2019.00121

30. *Hochreiter S., Schmidhuber J.* Long short-term memory // Neural computation. 1997. V. 9. Iss. 8. P. 1735–1780.

31. *Hoeffding W.* Probability inequalities for sums of bounded random variables // J. American Statistical Association. 1963. V. 58. Iss. 301. P. 13–30. https://doi.org/10.1080/01621459.1963.10500830

AUTONOMOUS SPACECRAFT CONTROL IN THE SOLAR GRAVITATIONAL LENS' FOCUS VIA REINFORCEMENT LEARNING

© 2025 M. G. Shirobokov*, K. R. Korneev, D. G. Perepukhov

Keldysh Institute of Applied Mathematics, Miusskaya Pl., 4, Moscow, Russia

**e-mail: shirobokov@keldysh.ru*

The problem of autonomous control of the translational motion of the spacecraft in the vicinity of the focus of the gravitational lens of the Sun is formulated. The problem is solved by a reinforcement machine learning method using contemporary stochastic numerical methods. The costs of the characteristic velocity for targeting the focal line of a remote extended source, the final accuracy of targeting and the quality of the control function are investigated. The results of the study are given for various forms of state and observation: 1) position and velocity, 2) noisy position and velocity, 3) image of the Einstein ring. The efficiency of control strategies when using recurrent layers and fully connected layers with an input in the form of a measurement stack is compared. The training of control models accounting for execution errors of maneuvers is also being explored.