XLIX Академические чтения по космонавтике
Москва, 28–31 января 2025 г.

Поддержание движения в окрестности гало-орбит системы Земля-Луна с использованием методов обучения с подкреплением

Широбоков М.Г., Перепухов Д.Г., Забара И.Д.

Институт прикладной математики им. М.В. Келдыша РАН

Мотивация работы

Высокий интерес к окололунному пространству и Луне вызван планами

1. строительства орбитальной окололунной станции
2. развертывания спутниковой навигационной сети
3. добычи полезных ресурсов на Луне
4. проведения научных исследований
5. демонстрации и тестирования новых технологий частными фирмами

Для успешной реализации этих планов важными задачами остаются обеспечение надежности миссий и парирование нештатных ситуаций.

Решение этих задач требует повышения автономности и адаптивности аппаратов, поскольку окололунное пространство — это сложная динамическая среда, где задержки сигналов затрудняют вмешательство с Земли, а аппараты должны оперативно реагировать на возмущения, сбои и внешние воздействия.

Методы синтеза автономного и адаптивного управления

На основе теории устойчивости:

- Методы на базе прямого метода Ляпунова
- Адаптивные регуляторы (model reference adaptive control)
- Робастное или гарантирующее управление (H_∞)
- Скользящее управление (sliding mode control)

На основе методов оптимизации:

- Методы прогнозирующего управления (model predictive control)
- Динамическое программирование Беллмана
- Нейросетевые адаптивные регуляторы
- Управление на базе обучения с подкреплением

На основе логики, эвристик и правил:

- Нечеткое управление
- Экспертные системы

Цель работы

Цель работы – разработка методики поддержания движения в окрестности гало-орбит, связанных с лунными точками либрации L_1 и L_2, на базе методов машинного обучения с подкреплением

Целесообразно учесть существование эффективных и фактически применяемых 50 лет методов поддержания либрационных орбит: метод нацеливание на точки орбиты (target point approach) и метод мод Флоке (Floquet mode approach)

Данный доклад посвящен такой методике и предоставляет результаты построения алгоритма управления в «простой» модели движения – круговой ограниченной задаче двух тел – и без неопределенности, чтобы дать представление о возможностях алгоритма в идеальных условиях

Постановка задачи

В рамках модели круговой ограниченной задачи трех тел рассмотрим гало-орбиту, связанную с точкой либрации L_1 или L_2 системы Земля–Луна. Точки орбиты в фазовом пространстве параметризуем как $\boldsymbol{x}_{\text{ref}}(\varphi)$, $\varphi \in [0, 2\pi]$. Период орбиты обозначим P.

Введем окрестность орбиты

$$\Omega_0 = \{\boldsymbol{x} = [\boldsymbol{r}, \boldsymbol{v}] : |\boldsymbol{r} - \boldsymbol{r}_{\text{ref}}(\varphi^*)| \le R_0, |\boldsymbol{v} - \boldsymbol{v}_{\text{ref}}(\varphi^*)| \le V_0,\ \varphi^* = \arg\min_{\varphi} |\boldsymbol{x} - \boldsymbol{x}_{\text{ref}}(\varphi)|\}.$$

В начальный момент времени t_0 фазовый вектор аппарата $\boldsymbol{x}_0 \in \Omega_0$ лежит в окрестности орбиты. Требуется построить функцию управления $\pi = \pi(\boldsymbol{x})$, которая выдает импульсы скорости по состоянию аппарата $\boldsymbol{x}$ в каждый P/k момент времени и поддерживает движение аппарата в окрестности Ω_0 орбиты.

Что такое обучение с подкреплением?

Агент взаимодействует со *средой*, получает от нее за свои действия скалярный сигнал – *вознаграждение* (подкрепление), и корректирует свое поведение так, чтобы добиться максимального вознаграждения.

Агент не знает, каково оптимальное поведение, и может даже не знать модель среды (модель движения) и вид функции вознаграждения. Он учится оптимальному поведению исключительно на опыте взаимодействия с ней, регистрируя цепочки состояние-действие-вознаграждение-состояние-действие-вознаграждение...

Стратегия поведения – это отображение из состояния в действие. Это отображение содержит параметры, значения которых нужно настроить так, чтобы поведение было оптимальным. Настройка значений параметров на основе опыта называется *обучением*.

Пример. Агент – управляющее программное обеспечение на космическом аппарате. Среда – космический аппарат. Действие – импульсы скорости. Вознаграждение – приближение к цели и экономия топлива. Стратегия – функция из состояния (положение/скорость) в действие (импульсы скорости).

Сведение к задаче обучения с подкреплением (1)

Общая методика: Широбоков М.Г. Методика построения управления космическими аппаратами с использованием методов обучения с подкреплением // Космические исследования. 2024. Т. 62, №5. С. 498–515.

1. Состояние – положение и скорость аппарата $\boldsymbol{x} = [\boldsymbol{r}, \boldsymbol{v}]$.

2. Область начальных состояний – $\Omega_0(R_0, V_0)$, где $R_0 = 100$ км, $V_0 = 1$ м/с. Распределение на Ω_0 – равномерное.

3. Шаг моделирования – интегрирование на интервале P/k, где k задано.

4. Функция вознаграждения:

$$r = -\alpha|\boldsymbol{x} - \boldsymbol{x}_{\text{ref}}(\varphi^*)|, \ \varphi^* = \arg\min_{\varphi} |\boldsymbol{x} - \boldsymbol{x}_{\text{ref}}(\varphi)|$$

Сведение к задаче обучения с подкреплением (2)

5. Модель восприятия:

$$\boldsymbol{o}(\boldsymbol{x}) = [\beta(\boldsymbol{x} - \boldsymbol{x}_{\text{ref}}(\varphi^*)),\, \cos\varphi^*,\, \sin\varphi^*] \in \mathbb{R}^8,\ \ \varphi^* = \arg\min_{\varphi} |\boldsymbol{x} - \boldsymbol{x}_{\text{ref}}(\varphi)|.$$

6. Параметрическая модель управления:

$$\pi(\boldsymbol{o}) = \pi_{FMA}(\boldsymbol{o}) + \gamma \cdot (\boldsymbol{A}_2\sigma(\boldsymbol{A}_1\boldsymbol{o} + \boldsymbol{b}_1) + \boldsymbol{b}_2),\ \ \pi_{FMA}(\boldsymbol{o}) = -\boldsymbol{e}_v \frac{\boldsymbol{e}^T(\boldsymbol{x} - \boldsymbol{x}_{\text{ref}}(\varphi^*))}{|\boldsymbol{e}_v|^2}.$$

Функции $\pi(\boldsymbol{o})$ и $\pi_{FMA}(\boldsymbol{o})$ отображают наблюдение $\boldsymbol{o}$ в импульсы скорости $\Delta\boldsymbol{v}$. Функция $\pi_{FMA}(\boldsymbol{o})$ отображает в минимальный импульс скорости, который устраняет проекцию относительного фазового вектора на неустойчивое направление $\boldsymbol{e} = [\boldsymbol{e}_r, \boldsymbol{e}_v]$ (собственный вектор, соответствующий максимальному по модулю собственному значению матрицы монодромии, соответствующей точке φ^*).

Алгоритм обучения с подкреплением

Алгоритм обучения: Proximal Policy Optimization (PPO)

Используемые библиотеки: `kiam_rl`, `stable-baselines3`

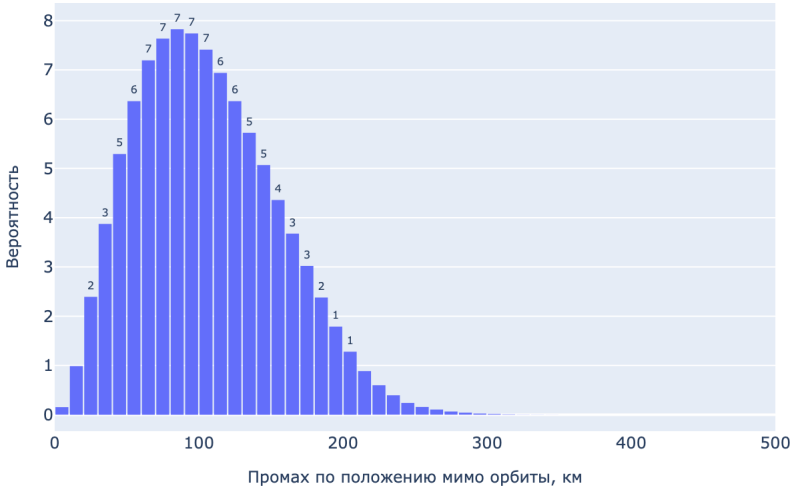
Общая схема расчетов:

1. Параметры модели управления фиксируются.
2. Модель управления испытывается в серии испытаний Монте-Карло, записываются все цепочки состояние-действие-вознаграждение.
3. На основании полученного опыта алгоритм PPO корректирует параметры модели управления. Возвращение в пункт 1.

Одно испытание – один шаг моделирования. После шага моделирования состояние инициализируется заново.

Управление методом мод Флоке

Гистограмма промаха по положению во время поддержания гало-орбиты L_1 с $A_z = 35365$ км, $P = 12$ дней, $\mu_1 = 741.21$



Модель управления:
Только FMA
Импульс каждые $1/4$ витка

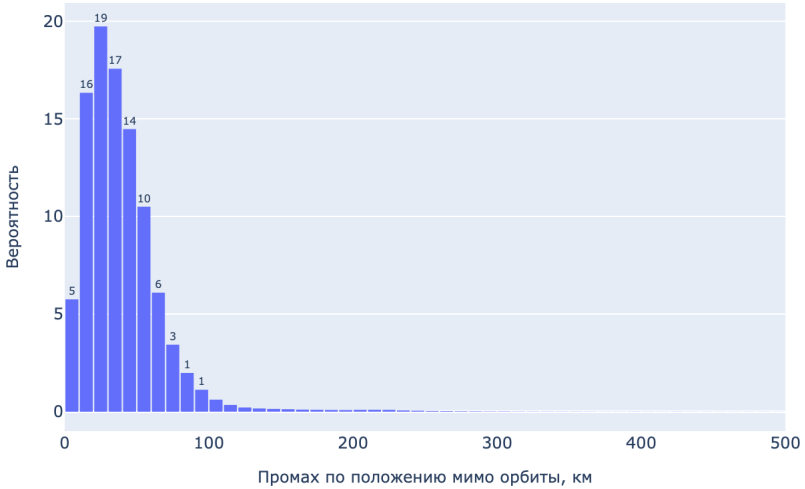
152019 испытаний на одном витке (точность оценки вероятностей 0.5%)

Квартили затрат хар. скорости за период (м/с):
0.00 0.09 0.19 0.35 1.40

Вероятность промаха более чем на 1000 км: 0.0%

Управление без метода мод Флоке

Гистограмма промаха по положению во время поддержания гало-орбиты L_1 с $A_z = 35365$ км, $P = 12$ дней, $\mu_1 = 741.21$



Модель управления:

$$(8) \xrightarrow{\text{L,th}} (16) \xrightarrow{\text{L}} (3)$$

Импульс каждые $1/4$ витка

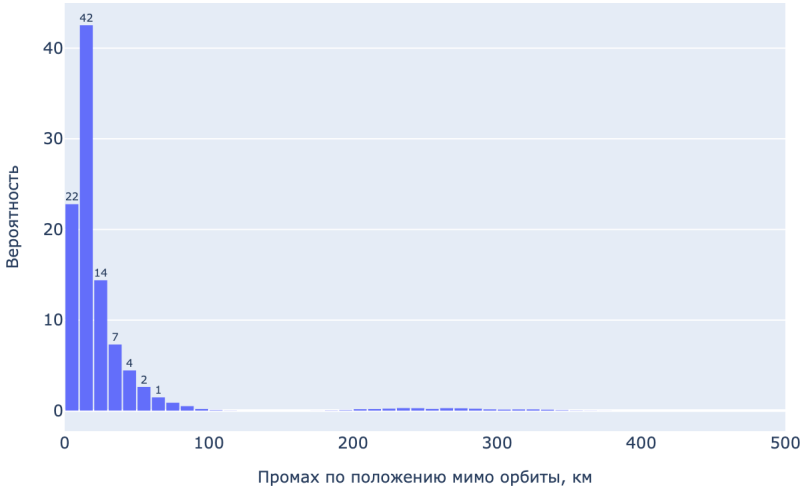
152019 испытаний на одном витке (точность оценки вероятностей 0.5%)

Квартили затрат хар. скорости за период (м/с):
0.15 1.05 1.30 1.58 4.69

Вероятность промаха более чем на 1000 км: 0.25%

Управление с методом мод Флоке

Гистограмма промаха по положению во время поддержания гало-орбиты L_1 с $A_z = 35365$ км, $P = 12$ дней, $\mu_1 = 741.21$



Модель управления:

FMA + (8) $\xrightarrow{\text{L,th}}$ (16) $\xrightarrow{\text{L}}$ (3)

Импульс каждые $1/4$ витка

152019 испытаний на одном витке (точность оценки вероятностей 0.5%)

Квартили затрат хар. скорости за период (м/с):
0.12 0.86 1.07 1.29 3.14

Вероятность промаха более чем на 1000 км: 0.0%

Заключение

Разработана методика поддержания движения в окрестности гало-орбит, связанных с лунными точками либрации L_1 и L_2, на базе методов машинного обучения с подкреплением, *учитывающая классический алгоритм поддержания методом мод Флоке*

Для неустойчивой L_1-гало-орбиты оценено распределение промаха по положению, затраты характеристической скорости, а также вероятность большого промаха в случаях применения управления 1) на основе только метода мод Флоке, 2) на основе нейросетевой модели без метода мод Флоке, 3) на основе метода мод Флоке и корректирующей его нейросетевой модели

Управление на основе только метода Флоке имеет низкие затраты скорости, но низкую точность поддержания. Управление на основе только нейросетевой модели существенно точнее поддерживает орбиту, но требует больше затрат на поддержание. Гибридное управление сохраняет высокую точность поддержания, при этом понижает затраты по сравнению с предыдущим случаем

Работа поддержана грантом Российского научного фонда (проект №24-71-00032).