13-я мультиконференция по проблемам управления
Математическая теория управления и ее приложения (МТУиП-2020)
6—8 октября 2020 г., Санкт-Петербург, Россия

Методы обучения с подкреплением в задачах управления движением космических аппаратов

Широбоков М.Г.

Институт прикладной математики им. М.В. Келдыша РАН

Цель доклада

Цель доклада – рассказать о новых задачах механики космического полета, требующих принципиально новые подходы к проектированию оптимального управления, и о том, как методы обучения с подкреплением помогают их решать и какие результаты были при этом получены

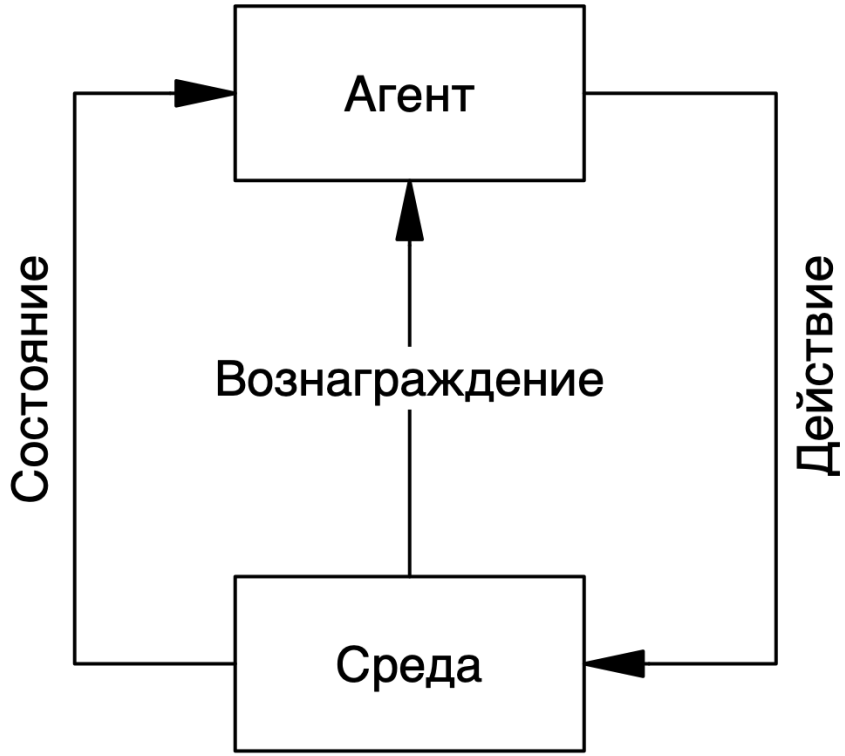
Новые задачи механики космического полета, которые требуют новых подходов к решению

- Посадка космического аппарата на поверхность небесного тела в условиях неопределенности рельефа, навигационной неопределенности, ошибок исполнения управления
- Адаптивное управление угловым движением аппарата при стыковке с объектом с неопределенными инерционными характеристиками
- Проектирование замкнутого контура управления для межпланетных траекторий с малой тягой при наличии сильных возмущений, навигационных ошибок и ошибок исполнения управления

Сравнение обучения с подкреплением со стандартной теорией оптимального управления в механике космического полета

Оптимальное управление	Обучение с подкреплением
Дает одну траекторию с оптимальным управлением без обратной связи	Дает квазиоптимальный закон управления с обратной связью
Требует закон поддержания траектории в условиях возмущений	Не требует поддержания траектории, управление уже учитывает возможные возмущения
Динамика представляется с помощью дифференциальных уравнений	Нет ограничений на представление динамики
Естественные функционал и ограничения	Функционал и ограничения выражаются в единой функции, называемой вознаграждением
Детерминированная оптимизация, предположение о детерминистских моделях	Стохастическая оптимизация, модели естественным образом считаются стохастическими

Общая постановка задачи обучения с подкреплением



Общая постановка задачи обучения с подкреплением

Дан марковский процесс принятия решений $(\mathcal{S}, \mathcal{A}, P, d_R, d_0, \gamma)$, где $\mathcal{S}$ – пространство состояний, $\mathcal{A}$ – пространство действий, функция перехода $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$

$$P(s, a, s') \coloneqq \Pr(S_{t+1} = s' | S_t = s, A_t = a)$$

функция d_R – условное распределение вознаграждений при заданных S_t, A_t и S_{t+1}, функция d_0 – начальное распределение состояний и $\gamma \in [0, 1]$ – параметр. Ставится задача поиска стратегии

$$\pi(s, a) \coloneqq \Pr(A_t = a | S_t = s)$$

максимизирующей функционал

$$J(\pi) \coloneqq \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| \pi\right]$$

Функции ценности и уравнения Беллмана

Функция ценности состояния: $v^{\pi}(s) := \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| S_0 = s, \pi\right]$

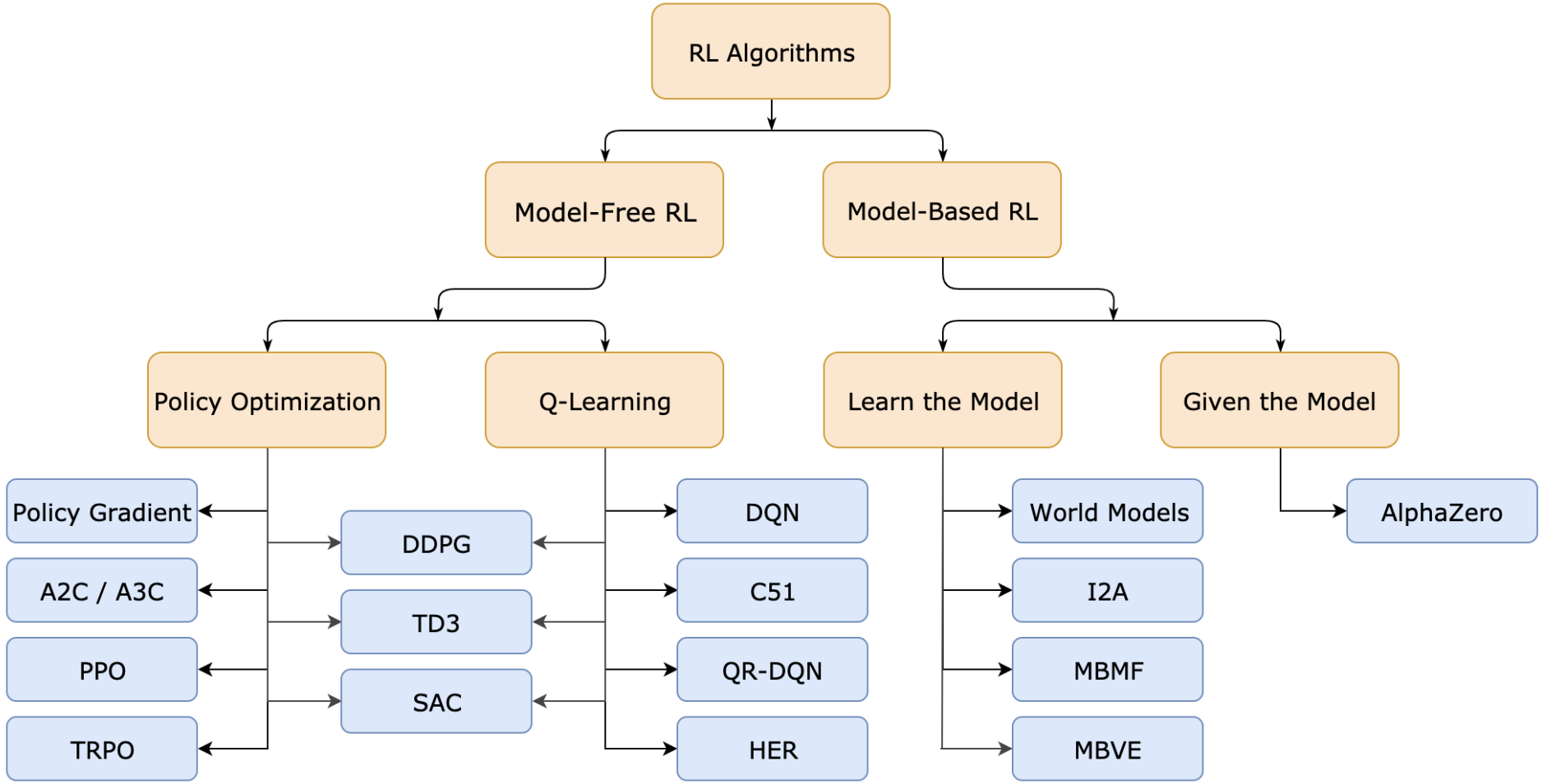
Функция ценности действия: $q^{\pi}(s,a) := \mathbf{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \middle| S_0 = s, A_0 = a, \pi\right]$

Уравнения оптимальности Беллмана:

$$v^*(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s,a,s') \left[R(s,a) + \gamma v^*(s')\right]$$

$$q^*(s,a) = \sum_{s' \in \mathcal{S}} P(s,a,s') \left[R(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*(s',a')\right]$$

Методы обучения с подкреплением



Credit: https://medium.com/analytics-vidhya/reinforcement-learning-machines-learning-by-interacting-with-the-world-64e5862dbf19

Deep Deterministic Policy Gradient (DDPG) и Twin Delay DDPG (TD3)

Методы DDPG и TD3 ищут оптимальную стратегию как решение уравнения оптимальности Беллмана

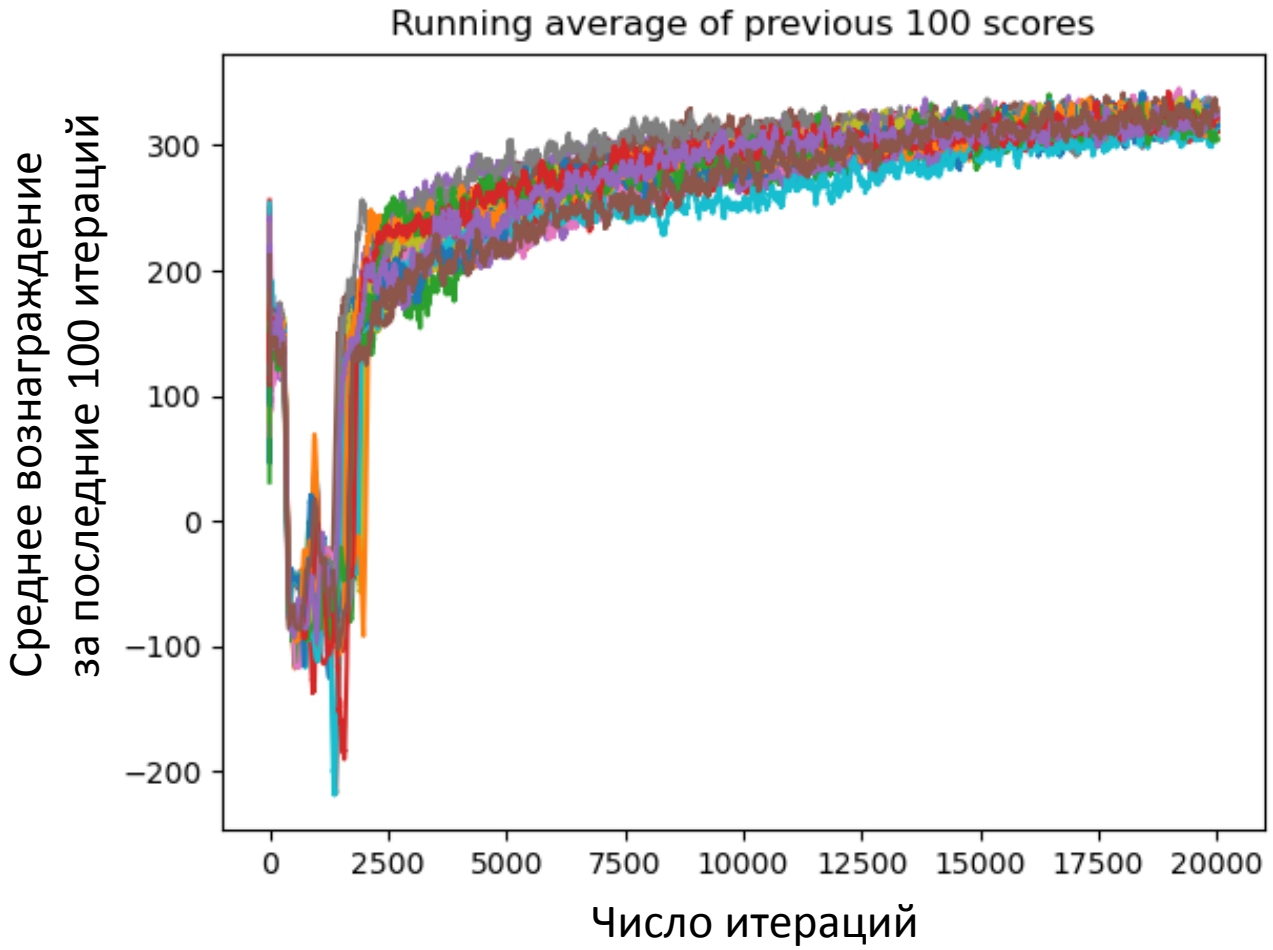
$$q^*(s,a) = \sum_{s' \in \mathcal{S}} P(s,a,s') \left[R(s,a) + \gamma \max_{a' \in \mathcal{A}} q^*(s',a') \right]$$

Стратегия и Q-функция параметризуются. Стратегия детерминированная. Вообще в методах Q-обучения в процессе оптимизации минимизируется невязка между левой и правой частью уравнений Беллмана. Методы DDPG и TD3 – это просто варианты борьбы с возникающими в непрерывном случае проблемами

Proximal Policy Optimization (PPO)

- Этот метод не использует уравнение оптимальности Беллмана и не пытается его решить
- Метод не использует понятие Q-функции
- PPO – это **метод доверительных областей** среди методов обучения с подкреплением и является модификацией похожего метода Trust Region Policy Optimization (TRPO)
- В отличие от DDPG и TD3 этот метод сначала копит опыт взаимодействия агента со средой, а потом на основании этого полученного опыта производит коррекцию параметров стратегии, причем коррекция эта – безопасная

Пример решения двухточечной краевой задачи в модели двух тел (задача Ламберта)



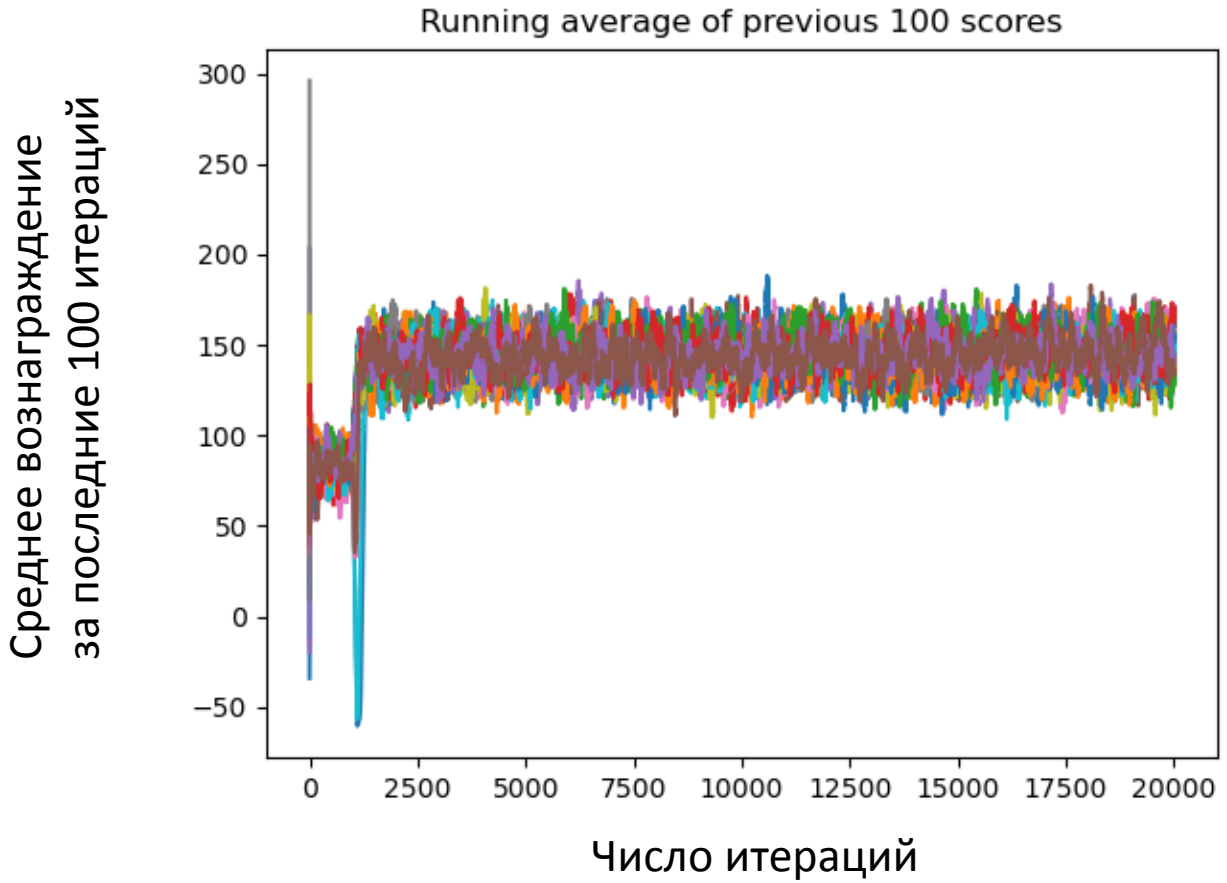
Даны краевые условия – начальное и конечное положения аппарата $\boldsymbol{r}_1$ и $\boldsymbol{r}_2$, соответственно и моменты времени t_1 и t_2

Найти скорость $\boldsymbol{v}_1$, переводящую аппарат из $\boldsymbol{r}_1$ в $\boldsymbol{r}_2$ на интервале времени $[t_1, t_2]$ в задаче двух тел

Эта задача может быть сформулирована как задача обучения с подкреплением, где состоянием является $\boldsymbol{r}(t_2)$, действием – $\boldsymbol{v}(t_1)$, а вознаграждение рассчитывается по формуле:

$$R = -100 \log_{10} |\boldsymbol{r}(t_2) - \boldsymbol{r}_2|$$

Пример решения двухточечной краевой задачи в модели двух тел (задача Ламберта)



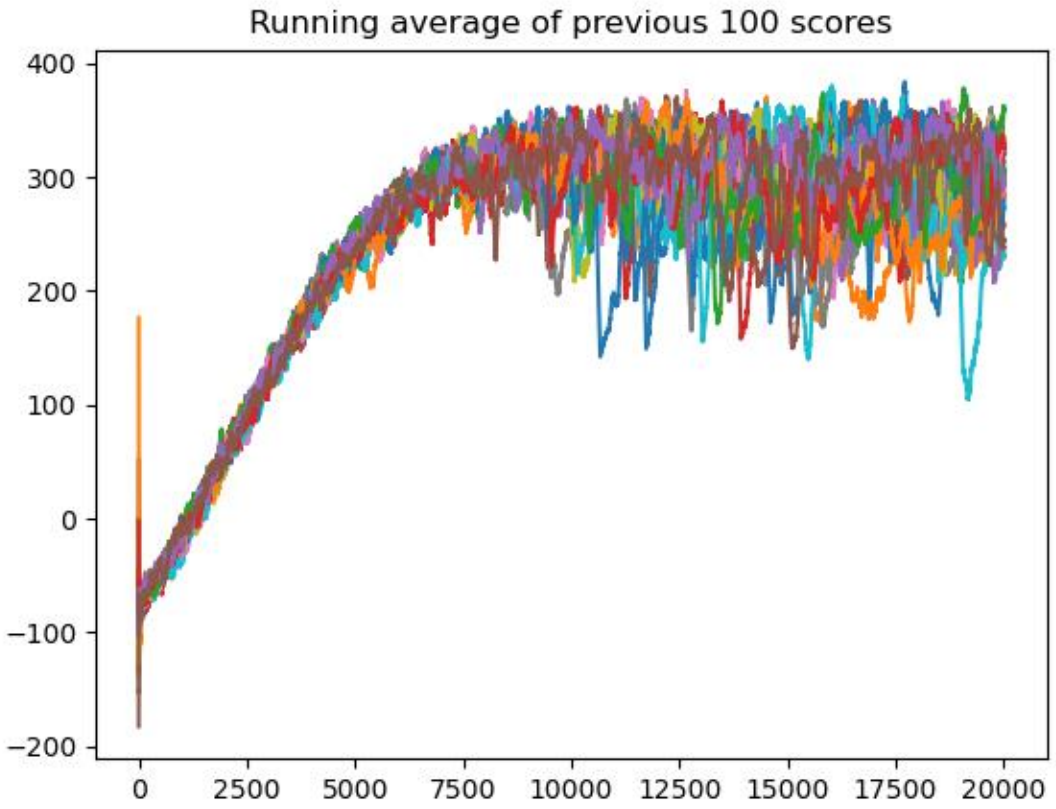
Даны краевые условия – начальное и конечное положения аппарата $\boldsymbol{r}_1$ и $\boldsymbol{r}_2$, соответственно и моменты времени t_1 и t_2

Найти скорость $\boldsymbol{v}_1$, переводящую аппарат из $\boldsymbol{r}_1$ в $\boldsymbol{r}_2$ на интервале времени $[t_1, t_2]$ в задаче двух тел

Эта задача может быть сформулирована как задача обучения с подкреплением, где состоянием является $\boldsymbol{r}(t_2)$, действием – $\boldsymbol{v}(t_1)$, а вознаграждение рассчитывается по формуле:

$$R = -100 \log_{10} |\boldsymbol{r}(t_2) - \boldsymbol{r}_2|$$

Пример решения двухточечной краевой задачи в модели двух тел (задача Ламберта)



Даны краевые условия – начальное и конечное положения аппарата $\boldsymbol{r}_1$ и $\boldsymbol{r}_2$, соответственно и моменты времени t_1 и t_2

Найти скорость $\boldsymbol{v}_1$, переводящую аппарат из $\boldsymbol{r}_1$ в $\boldsymbol{r}_2$ на интервале времени $[t_1, t_2]$ в задаче двух тел

Эта задача может быть сформулирована как задача обучения с подкреплением, где состоянием является $\boldsymbol{r}(t_2)$, действием – $\boldsymbol{v}(t_1)$, а вознаграждение рассчитывается по формуле:

$$R = -100 \log_{10} |\boldsymbol{r}(t_2) - \boldsymbol{r}_2|$$

Когда метод обучения с подкреплением оказывается лучше традиционного

Недавняя работа

Gaudet B., Linares R., Furfaro R. Adaptive guidance and integrated navigation with reinforcement meta-learning. Acta Astronautica, 2020. Vol. 169. pp. 180-190. doi: 10.1016/j.actaastro.2020.01.007

посвящена адаптивной системе управления и наведения при посадке на Марс, построенную методами обучения с подкреплением. Управление, построенное методами обучения с подкреплением оказалось (по сравнению с традиционным алгоритмом наведения с обратной связью): устойчивым к частым отказам двигателя, вариации удельного импульса и ошибкам определения состояния аппарата

Заключение

- В последнее время появляются новые задачи механики космического полета, которые требуют новых подходов к решению – особенностью этих задач является стохастичность постановок и требование построить надежное управление
- Методы обучения с подкреплением для задач управления (непрерывные множества состояний и действий) значительно развились за последние 5 лет, среди них в механике полета особенно часто применяются PPO, DDPG и TD3
- Эти методы не требуют модель среды, не требуют градиентов функционала и ограничений по переменным модели, долго работают даже на простейших задачах, но позволяют получить надежное управление в очень сложных и реалистичных постановках