Основанные на модели методы обучения с подкреплением для задач управления динамическими объектами

Лебедь Федор Сергеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: fedorlebed.cs@yandex.ru

Научный руководитель *— Кропотов Дмитрий Александрович*

В классической постановке задача обучения с подкреплением [3] записывается следующим образом:

$$\begin{cases} R(\pi) = \mathbb{E} \sum\limits_{t=0}^{\infty} \gamma^t r(s_t, a_t) \longrightarrow \min\limits_{\pi \in \Pi}, \\ a_t \sim \pi(s_t), \\ s_{t+1} \sim p(s_t, a_t), \\ s_0 \sim p_0(s_0), \end{cases} \tag{1}$$

где $\{s_t\}$ – последовательность состояний системы, $\{a_t\}$ – последовательность действий агента, π – обучаемая стратегия агента, $\{p, p_0\}$ – динамика среды, в котором действует агент, r – штрафная функция, а $\gamma \in (0, 1)$ – дисконтирующий фактор.

В такой постановке динамика среды $\{p, p_0\}$ считается неизвестной, состояния $\{s_t\}$ и действия $\{a_t\}$ могут быть как дискретными, так и непрерывными, а штрафная функция r может быть сколь угодно "плохой". Вследствие чего методы решения задачи (1) страдают от неустойчивости, необходимости тщательного подбора гиперпараметров и большого числа взаимодействий со средой.

В данной работе предлагается ввести на задачу (1) следующие ограничения:

$$\begin{cases} s \in \mathbb{R}^n, \\ a \in \mathbb{R}^m, \\ r(s, a) \in C^2(\mathbb{R}^n \times \mathbb{R}^m) \text{ – } \textit{детерминированная функция}, \\ p(s, a) \in C^2(\mathbb{R}^n \times \mathbb{R}^m) \text{ – } \textit{детерминированная функция}, \\ \pi(s), \in C^1(\mathbb{R}^n \times \mathbb{R}^m) \text{ – } \textit{детерминированная функция}. \end{cases} \tag{2}$$

В такой постановке "оптимальное" управление для фиксированного начального состояния s_0, оцененной динамики системы $\hat{p}(s, a) \approx$

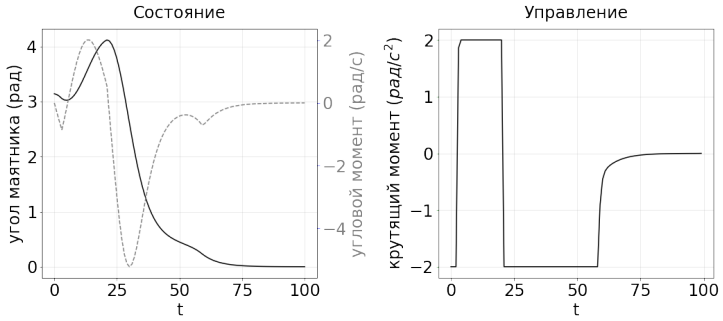
$p(s,a)$ и оцененной функции Беллмана $\hat{V}(s)$ [2] можно находить методом iLQR [1], примененным к следующей задаче:

$$\begin{cases} \sum\limits_{t=0}^{T} \gamma^t r(s_t, a_t) + \gamma^{T+1} \hat{V}(s_{T+1}) \longrightarrow \min\limits_{\{a_0 \dots a_T\}}, \\ s_{t+1} = \hat{p}(s_t, a_t). \end{cases} \tag{3}$$

На основании данных, полученных в ходе решений задачи (3) предлагается обучать стратегию агента π, а также улучшать оценки динамики среды $\hat{p}$ и функции Беллмана $\hat{V}(s)$.

В данной работе исследуется вопрос эффективности вышеописанного подхода в смысле количества взаимодействий агента со средой во время обучения, а также актуальность предложенного метода в некоторых подклассах задачи (1) с ограничениями вида (2).

Иллюстрации



Оптимальное решение для задачи закидывания маятника наверх.

Литература

1. J. Chen, W. Zhan, and M. Tomizuka, “Autonomous driving motion plan- ning with constrained iterative LQR,” IEEE Transactions on Intelligent Vehicles, vol. 4, no. 2, pp. 244–254, 2019.
2. Dixit, Avinash K. (1990). Optimization in Economic Theory (2nd ed.). Oxford University Press. p. 164.
3. Kaelbling, Leslie P.; Littman, Michael L.; Moore, Andrew W. (1996). “Reinforcement Learning: A Survey”. Journal of Artificial Intelligence Research. 4: 237–285.