УДК 004.81

М.З. ЗГУРОВСЬКИЙ
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна,
e-mail: *mzz@kpi.ua.*

П.О. КАСЬЯНОВ
Навчально-науковий комплекс «Інститут прикладного системного аналізу» Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» МОН України та НАН України, Київ, Україна,
e-mail: *kasyanov@i.ua*.

Л.Б. ЛЕВЕНЧУК
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна,
e-mail: *lusi.levenchuk@gmail.com*.

ФОРМАЛІЗАЦІЯ МЕТОДІВ ПОБУДОВИ АВТОНОМНИХ СИСТЕМ ШТУЧНОГО ІНТЕЛЕКТУ

Анотація. Розв'язано задачу формалізації побудови автономних систем штучного інтелекту, математичні моделі яких можуть бути складними або неідентифікованими. За допомогою методу послідовних наближень для Q-функцій винагород розроблено методологію побудови наближених за заданою точністю ε-оптимальних стратегій. Результати дають змогу визначити класи (зокрема, подвійного призначення), для яких можна на сучасному рівні математичної строгості обґрунтовано будувати оптимальні та ε-оптимальні стратегії навіть у випадках, коли моделі ідентифікуються, але обчислювальна складність стандартних алгоритмів динамічного програмування може не бути строго поліноміальною.

Ключові слова: автономні системи штучного інтелекту, марковські процеси прийняття рішень, ε-оптимальні стратегії.

ВСТУП

Автономні системи штучного інтелекту (АСШІ) — це системи, які можуть виконувати покладені на них завдання без зовнішнього втручання людини. Наведемо неповний перелік автономних систем штучного інтелекту, а саме [1–30]:

— автомобілі, літаки та інші рухомі об'єкти, якими керує автопілот: вони самостійно приймають рішення про траєкторію і програму руху без участі людини;

— літальні або плавальні дрони: це автономні апарати, які можуть виконувати різноманітні завдання, а саме доставлення певних речей у певні точки простору, або дистанційно сканувати стан окремих територій;

— інтернет речей (IoT): це системи, що містять різноманітні датчики та «розумні» пристрої, які можуть збирати дані про навколишнє середовище та виконувати різноманітні дії відповідно до покладених на них завдань;

— системи самообслуговування: це системи, які дають змогу клієнтам купувати товари або отримувати інші послуги, не користуючись порадами співробітників торговельних центрів;

— автоматизовані медичні системи, які дають змогу діагностувати та надавати поради щодо лікування хворих, не залучаючи лікарів;

— роботизовані фінансові системи, які можуть автоматично розподіляти активи та керувати ризиками на фінансових ринках;

— автоматизовані виробничі системи, які містять різноманітні автомати та роботи, що можуть автоматично виконувати завдання виробничого процесу.

Отже, маємо багато систем, які можуть виконувати покладені на них функції без участі людини.

Операційні середовища, для яких створюються і в яких мають діяти нинішні і майбутні АСШІ, швидко розвиваються. Особливої актуальності набувають проблеми функціонування автономних систем подвійного призначення, які мають протидіяти аналогічним АСШІ в конкурентних умовах [30]. Ефективність цих систем залежатиме від їхніх можливостей забезпечити своєчасну інтелектуальну та надійну підтримку прийняття рішень за швидкозмінних операційних умов та неповноти інформації стосовно стратегії і програми дій супротивників. Робота таких систем також має базуватися на сучасних методах та інструментах керування ризиками.

ПОБУДОВА АСШІ В ПАРАДИГМІ МАРКОВСЬКИХ ПРОЦЕСІВ ПРИЙНЯТТЯ РІШЕНЬ

Марковські процеси прийняття рішень (МППР) — це модель послідовного прийняття рішень в умовах невизначеності, яка ґрунтується на теорії послідовних статистичних проблем прийняття рішень, розробленій під час Другої світової війни, і вивчається більше сімдесяти років у контексті дослідження операцій та суміжних галузей. Останнім часом МППР широко застосовуються в задачах машинного навчання, оскільки вони є стандартною структурою для моделювання, розроблення та аналізу алгоритмів навчання з підкріпленням [7]. МППР моделюють проблеми, коли агент взаємодіє з навколишнім середовищем з плином часу. На поточний стан навколишнього середовища можуть впливати як дії агента, так і фактори, що не перебувають під контролем агента.

Агент має конкретну мету (наприклад, виявити приховану ціль, перейти до певного місця) і повинен визначити взаємодії з навколишнім середовищем (наприклад, де шукати, в якому напрямку рухатися), щоб найкращим чином досягти своєї мети. Метою агента зазвичай є пошук політики (наприклад, відображення, яке діє із простору станів в простір дій), що мінімізує очікувані загальні витрати на горизонті планування [6]. Якщо динаміка МППР (тобто ймовірності переходу) повністю відома, тоді оптимальні політики за цими критеріями можуть бути обчислені за допомогою стандартних методів [6].

В іншому випадку для пошуку політик можуть бути використані апроксимаційні імітаційні версії таких методів, а для цього потрібно використовувати нейронні мережі. Навчання з підкріпленням [7] застосовують для розроблення та аналізу різних типів алгоритмів для МППР.

Фахівцями компанії штучного інтелекту DeepMind була сформульована гіпотеза [11]: «інтелект і пов'язані з ним здібності можна розуміти як забезпечення максимізації винагороди агентом, що діє в своєму середовищі». Інакше кажучи, знаючи винагороди для кожної пари (стан, дія), теоретично завжди можна формалізувати проблему побудови АСШІ як задачу навчання з підкріпленням, яка зазвичай формалізується за допомогою частково спостережуваних марковських процесів прийняття рішень, де відомими є простори станів, допустимих керувань та винагороди.

Сама модель зазвичай є невідомою та ідентифікується одночасно із пошуком апроксимацій оптимальних дій. Для розроблення обґрунтованих на сучасному рівні математичної строгості алгоритмів підтримки прийняття рішень потрібно, щоб проблема покрокового прийняття рішень була сформульована математично коректно. Ключовою вимогою до математично формалізованої задачі є те, що вона повинна бути розв'язною.

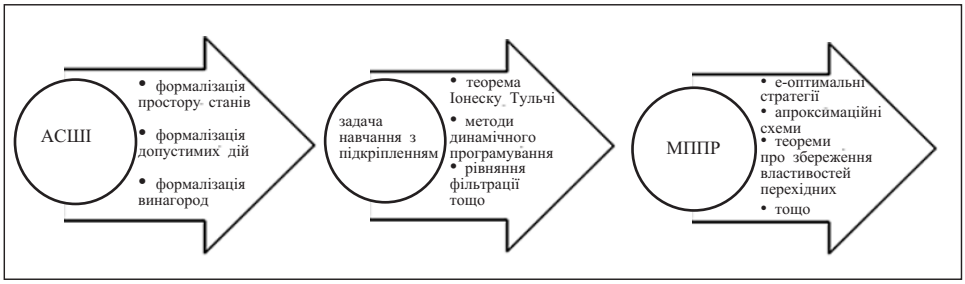

Рис. 1. Формалізація побудови автономних систем штучного інтелекту

Для складних задач послідовного прийняття рішень часто нетривіально визначити розв'язність моделі [5]. Виникає необхідність розроблення математичниого апарату для розв'язання задач послідовного прийняття рішень з метою забезпечення надійного та робастного прийняття рішень у середовищах, з якими взаємодіють автономні системи, а також розвивати доказово ефективні алгоритми для обчислення рекомендацій щодо дій агента (рис. 1).

ПОШУК ε-ОПТИМАЛЬНИХ СТРАТЕГІЙ

Формалізуючи простори станів та дій, а також винагороди від різних дій у кожному стані, поставимо задачу побудови АСШІ як задачу навчання з підкріпленням [11]. Відповідно до [12, 13] формалізуємо цю задачу в загальному випадку частково спостережуваних МППР, де апроксимацію оптимальних стратегій визначитимемо за допомогою Q-функцій. Для того щоб окреслити межі застосовності класів таких проблем, спочатку уведемо допоміжні означення та твердження. Потім обґрунтуємо збіжність алгоритмів пошуку наближених стратегій у динаміці.

Для метричного простору $\mathbb{S}$ розглянемо Борелеву σ-алгебру $\mathcal{B}(\mathbb{S})$ і позначимо $\mathbb{P}(\mathbb{S})$ множину ймовірнісних мір $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$. Послідовність ймовірнісних мір $\{\mu^{(n)}\}_{n=1,2,\ldots}$ з $\mathbb{P}(\mathbb{S})$ слабко збігається до $\mu \in \mathbb{P}(\mathbb{S})$, якщо для будь-якої неперервної обмеженої функції f на просторі $\mathbb{S}$ має місце

$$\int_{\mathbb{S}} f(s)\, \mu^{(n)}(ds) \to \int_{\mathbb{S}} f(s)\, \mu(ds), \qquad n \to \infty.$$

Послідовність ймовірнісних мір $\{\mu^{(n)}\}_{n=1,2,\ldots}$ з множини $\mathbb{P}(\mathbb{S})$ збігається за варіацією до $\mu \in \mathbb{P}(\mathbb{S})$, якщо

$$\sup\left\{\int_{\mathbb{S}} f(s)\, \mu^{(n)}(ds) \to \int_{\mathbb{S}} f(s)\, \mu(ds) \colon f \colon \mathbb{S} \to [-1,1] \text{ є вимірною за Борелем}\right\} \to 0,$$

$$n \to \infty.$$

Зауважимо, що $\mathbb{P}(\mathbb{S})$ — сепарабельний метричний простір стосовно топології слабкої збіжності ймовірнісних мір, якщо $\mathbb{S}$ є сепарабельним метричним простором [14]. Для Борелевих просторів $\mathbb{S}_1$ та $\mathbb{S}_2$ стохастичним ядром $R(ds_1 | s_2)$ на $\mathbb{S}_1$ для заданого $\mathbb{S}_2$ є таке відображення $R(\cdot\,|\cdot) \colon \mathcal{B}(\mathbb{S}_1) \times \mathbb{S}_2 \to [0,1]$, що для будь-якого $s_2 \in \mathbb{S}_2$ $R(\cdot\,|s_2)$ є ймовірнісною мірою на $\mathbb{S}_1$ та $R(B|\cdot)$ є вимірною за Борелем функцією на $\mathbb{S}_2$ для будь-якої Борелевої множини $B \in \mathcal{B}(\mathbb{S}_1)$.

Стохастичне ядро $R(ds_1 | s_2)$ на $\mathbb{S}_1$ для заданого $\mathbb{S}_2$ визначає вимірне за Борелем відображення $s_2 \to R(\cdot\,|s_2)$ з $\mathbb{S}_2$ в метричний простір $\mathbb{P}(\mathbb{S}_1)$ з топологією слабкої збіжності ймовірністних мір. Стохастичне ядро $R(ds_1 | s_2)$ називають

слабко неперервним (неперервним за варіацією), якщо $R(\cdot|x^{(n)})$ збігається слабко (за варіацією) до $R(\cdot|x)$, коли $x^{(n)}$ збігається до x в $\mathbb{S}_2$.

Нехай $\mathbb{X}$, $\mathbb{Y}$ та $\mathbb{A}$ — Борелеві простори, $P(dx'|x,a)$ — стохастичне ядро на $\mathbb{X}$ для заданого $\mathbb{X}\times\mathbb{A}$, $Q(dy\,|\,a,x)$ — стохастичне ядро на $\mathbb{Y}$ для заданого $\mathbb{A}\times\mathbb{X}$, $Q_0(dy\,|\,x)$ — стохастичне ядро на $\mathbb{Y}$ для заданого $\mathbb{X}$, p — розподіл імовірностей на $\mathbb{X}$, $c\colon \mathbb{X}\times\mathbb{A}\to\overline{\mathbb{R}}=\mathbb{R}\cup\{+\infty\}$ — обмежена знизу Борелева функція на $\mathbb{X}\times\mathbb{A}$. Частково спостережувані МППР визначаються за допомогою $(\mathbb{X},\mathbb{Y},\mathbb{A},P,Q,r)$, де $\mathbb{X}$ — простір станів, $\mathbb{Y}$ — простір спостережень, $\mathbb{A}$ — множина дій, $P(dx'|x,a)$ — правило переходу для стану, $Q(dy|a,x)$ — спостережуване ядро, $r\colon \mathbb{X}\times\mathbb{A}\to\overline{\mathbb{R}}$ — однокрокова винагорода.

Частково спостережувані МППР еволюціонують за такими правилами:

• для $t=0$ початковий неспостережуваний стан x_0 має заданий апріорний розподіл p;

• початкове спостереження стану y_0 породжується відповідно до початкового ядра спостереження $Q_0(\cdot|x_0)$;

• у кожний момент часу $t=0,1,\ldots$, якщо стан системи $x_t\in\mathbb{X}$, агент вибирає дію $a_t\in\mathbb{A}$ та виплачується винагорода $r(x_t,a_t)$;

• система переходить в стан x_{t+1} за правилом переходу $P(\cdot|x_t,a_t)$, $t=0,1,\ldots$;

• спостереження $y_{t+1}\in\mathbb{Y}$ утворюється ядром $Q(\cdot|a_t,x_{t+1})$, $t=0,1,\ldots$, а спостереження y_0 в початковий момент часу породжується ядром спостереження $Q_0(\cdot|x_0)$.

Визначимо історії спостережень: $h_0:=(p,y_0)\in H_0$ та

$$h_t:=(p,\ y_0,\ a_0,\ldots,y_{t-1},\ a_{t-1},\ y_t)\in H_t \ \text{ для всіх } \ t=1,2,\ldots,$$

де $\mathbb{H}_0:=\mathbb{P}(\mathbb{X})\times\mathbb{Y}$ та $\mathbb{H}_t:=\mathbb{H}_{t-1}\times\mathbb{A}\times\mathbb{Y}$, якщо $t=1,2,\ldots$ Тоді стратегія π для частково спостережуваних МППР визначається як послідовність $\pi=\{\pi_t\}_{t=0,1,\ldots}$ стохастичних ядер π_t на $\mathbb{A}$ для заданого $\mathbb{H}_t$. Стратегія π називається нерандомізованою, якщо кожна ймовірнісна міра $\pi_t(\cdot|h_t)$ зосереджена в одній точці. Нерандомізована стратегія називається марковською, якщо всі рішення залежать тільки від поточного стану і часу. Марковська стратегія називається стаціонарною, якщо всі рішення залежать лише від поточного стану.

Позначимо Π множину всіх стратегій. З теореми Іонеску Тульчі [15, с. 140–141; 16, с. 178] випливає, що стратегія $\pi\in\Pi$, початковий розподіл $p\in\mathbb{P}(\mathbb{X})$, початкова дія a разом зі стохастичними ядрами P, Q та Q_0 формулюють єдину ймовірнісну міру $P^{\pi}_{p,a}$ на множині всіх траєкторій $\mathbb{H}_\infty:=\mathbb{P}(\mathbb{X})\times(\mathbb{Y}\times\mathbb{A})^\infty$ з σ-алгеброю, визначеною добутком Борелевих σ-алгебр на $\mathbb{P}(\mathbb{X})$, $\mathbb{Y}$ та $\mathbb{A}$. Математичне сподівання відносно цієї ймовірнісної міри позначимо $\mathbb{E}^{\pi}_{p}$. Задамо критерій якості. Для скінченного горизонту $T=0,1,\ldots$ визначимо очікувані загальні дисконтовані винагороди (Q-функції):

$$V^{\pi}_{T,\alpha}(p,a):=\mathbb{E}^{\pi}_{p,a}\sum_{t=0}^{T-1}\alpha^t r(x_t,a_t),\qquad \pi\in\Pi,\ p\in\mathbb{P}(\mathbb{X}),\ a\in\mathbb{A},\tag{1}$$

де $\alpha\geq 0$ — дисконтуючий множник, $V^{\pi}_{0,\alpha}(p,a)=0$. Оскільки Q визначає ядро спостережень, для Q-функції використовуватимемо позначення $V^{\pi}_{T,\alpha}$ — загальні дисконтовані винагороди.

У подальшому вважатимемо, що завжди виконується одна з таких умов.

Припущення (D): функція r обмежена зверху на $\mathbb{X}\times\mathbb{A}$ і $\alpha\in[0,1)$.

Припущення (P): функція r недодатна на $\mathbb{X}\times\mathbb{A}$ і $\alpha\in[0,1]$.

Для $T=\infty$ формула (1) визначає очікувані загальні дисконтовані винагороди з нескінченним горизонтом, які позначимо $V_\alpha^\pi(p,a)$. Для довільної функції $g^\pi(p,\,a)$, зокрема для $g^\pi(p,\,a)=V_{T,\alpha}^\pi(p,\,a)$ та $g^\pi(p,\,a)=V_\alpha^\pi(p,\,a)$, визначимо оптимальну винагороду $g(p,\,a):=\sup\limits_{\pi\in\Pi} g^\pi(p,\,a)$, $p\in\mathbb{P}(\mathbb{X})$, $a\in\mathbb{A}$, де Π — множина всіх стратегій.

Стратегія π називається оптимальною для відповідного критерію, якщо $g^\pi(p,a)=g(p,a)$ для всіх $p\in\mathbb{P}(\mathbb{X})$, $a\in\mathbb{A}$. Для $g^\pi=V_{T,\alpha}^\pi$ оптимальна стратегія називається дисконтовано-оптимальною зі скінченним горизонтом T, а для $g^\pi=V_\alpha^\pi$ — дисконтовано-оптимальною.

У цій статті використовуються аналогічні позначення також і для очікуваних загальних винагород і цільових значень марковських процесів прийняття рішень та повністю спостережуваних МППР. Для визначеності крім позначень $V_{T,\alpha}^\pi$, V_α^π, $V_{T,\alpha}$ та V_α, уведених для частково спостережуваних МППР, використовуватимемо позначення $v_{T,\alpha}^\pi$, v_α^π, $v_{T,\alpha}$, v_α і $\bar v_{T,\alpha}^\pi$, $\bar v_\alpha^\pi$, $\bar v_{T,\alpha}$, $\bar v_\alpha$ для МППР та повністю спостережуваних МППР відповідно.

Зауважимо, що функція r, визначена на $\mathbb{X}\times\mathbb{A}$ із значеннями в $\overline{\mathbb{R}}$, є sup-компактною, якщо множина $\{(x,a)\in\mathbb{X}\times\mathbb{A}: r(x,\,a)\ge\lambda\}$ є компактною для будь-якого скінченного λ. Функція r, визначена на $\mathbb{X}\times\mathbb{A}$ із значеннями в $\overline{\mathbb{R}}$, називається $\mathbb{K}$-sup-компактною на $\mathbb{X}\times\mathbb{A}$, якщо для будь-якої компактної множини $K\subseteq\mathbb{X}$ функція r є sup-компактною на $K\times\mathbb{A}$ [17].

Відповідно до [17, лема 2.5] обмежена зверху функція r є $\mathbb{K}$-sup-компактною на добутку метричних просторів $\mathbb{X}$ та $\mathbb{A}$ тоді і тільки тоді, коли вони задовольняють такі дві умови:

1) функція r є напівнеперервною зверху;

2) якщо послідовність $\{x^{(n)}\}_{n=1,2,\dots}$ із значеннями в $\mathbb{X}$ збігається та її границя x належить $\mathbb{X}$, то будь-яка послідовність $\{a^{(n)}\}_{n=1,2,\dots}$ з $a^{(n)}\in\mathbb{A}$, $n=1,2,\dots$, така, що послідовність $\{r(x^{(n)},a^{(n)})\}_{n=1,2,\dots}$ обмежена знизу, має граничну точку $a\in\mathbb{A}$.

Частково спостережувані МППР $(\mathbb{X},\mathbb{Y},\mathbb{A},P,Q,r)$ є повністю спостережуваними, якщо $\mathbb{Y}=\mathbb{X}$ та $Q(B\,|\,a,x)=Q(B\,|\,x)=\boldsymbol{I}\{x\in B\}$ для усіх $x\in\mathbb{X}$, $a\in\mathbb{A}$ і $B\in\mathcal{B}(\mathbb{X})$.

ЗАГАЛЬНІ УМОВИ ДЛЯ ІСНУВАННЯ ОПТИМАЛЬНИХ СТРАТЕГІЙ МППР

Визначимо загальні умови для існування оптимальних стратегій МППР [18], обгрунтованості рівнянь оптимальності і збіжності послідовних наближень для загальних очікуваних винагород. Сформулюємо ці умови.

Припущення (W^*) [18]:

(i) функція r є $\mathbb{K}$-sup-компактною на $\mathbb{X}\times\mathbb{A}$;

(ii) ядро $(x,\,a)\mapsto P(\cdot\,|\,x,\,a)$ є слабконеперервним на $\mathbb{X}\times\mathbb{A}$.

У цьому припущенні частково спостережувані МППР зводяться до повністю спостережуваних МППР [19–22, 24]. Для спрощення позначень ми іноді випускатимемо часовий параметр. Для заданого апостеріорного розподілу z стану x в момент часу $t = 0,1,\ldots$ і заданої дії a, вибраної за період t, позначимо $R(B \times C \mid z,\ a)$ сукупну ймовірність того, що в момент часу $(t+1)$ стан належить множині $B \in \mathcal{B}(\mathbb{X})$ та спостереження належить множині $C \in \mathcal{B}(\mathbb{Y})$:

$$R(B \times C \mid z,\ a) := \int_{\mathbb{X}} \Big| \int_{B} Q(C|a,x')P(dx'|x,\ a)\,z(dx), \tag{2}$$

$$B \in \mathcal{B}(\mathbb{X}),\ C \in \mathcal{B}(\mathbb{Y}),\ z \in \mathbb{P}(\mathbb{X}),\ a \in \mathbb{A},$$

де R — стохастичне ядро на $\mathbb{X} \times \mathbb{Y}$ для заданого $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$. Отже, визначимо ймовірність $R'(C \mid z, a)$ того, що спостереження в момент часу $t+1$ належить множині C:

$$R'(C \mid z,a) = \int_{\mathbb{X}} \int_{\mathbb{X}} Q(C \mid a,x')P(dx'|x,\ a)\,z(dx), \tag{3}$$

$$C \in \mathcal{B}(\mathbb{Y}),\ z \in \mathbb{P}(\mathbb{X}),\ a \in \mathbb{A},$$

де R' — стохастичне ядро на $\mathbb{Y}$ для заданого $\mathbb{P}(\mathbb{X}) \times \mathbb{A}$. Відповідно до [25] існує стохастичне ядро H на $\mathbb{X}$ заданого $\mathbb{P}(\mathbb{X}) \times \mathbb{A} \times \mathbb{Y}$ таке, що

$$R(B \times C \mid z,\ a) = \int_{C} H(B \mid z,a,y)R'(dy \mid z,a), \tag{4}$$

$$B \in \mathcal{B}(\mathbb{X}),\ C \in \mathcal{B}(\mathbb{Y}),\ z \in \mathbb{P}(\mathbb{X}),\ a \in \mathbb{A}.$$

Стохастичне ядро $H(\cdot \mid z, a, y)$ визначає вимірне відображення $\mathbb{H}: \mathbb{P}(\mathbb{X}) \times \mathbb{A} \times \mathbb{Y} \to \mathbb{P}(\mathbb{X})$, де $H(z, a, y)[\cdot] = H(\cdot \mid z, a, y)$. Для кожної пари $(z, a) \in \mathbb{P}(\mathbb{X}) \times \mathbb{A}$ відображення $\mathbb{H}(z, a, \cdot): \mathbb{Y} \to \mathbb{P}(\mathbb{X})$ визначається $R'(\cdot \mid z, a)$-м.н. однозначно [22, наслідок 7.27.1], [21, додаток 4.4]. Для апостеріорного розподілу $z_t \in \mathbb{P}(\mathbb{X})$, дії $a_t \in \mathbb{A}$ і спостереження $y_{t+1} \in \mathbb{Y}$ апостеріорний розподіл $z_{t+1} \in \mathbb{P}(\mathbb{X})$ задовольняє рівняння фільтрації:

$$z_{t+1} = H(z_t,\ a_t,\ y_{t+1}). \tag{5}$$

Зауважимо, що y_{t+1} — випадкова величина з розподілом $R'(\cdot \mid z_t, a_t)$, а права частина (5) відображає $(z_t, a_t) \in \mathbb{P}(\mathbb{X}) \times \mathbb{A}$ в $\mathbb{P}(\mathbb{P}(\mathbb{X}))$. Отже, z_{t+1} — випадкова величина зі значеннями в $\mathbb{P}(\mathbb{X})$, розподіл якої однозначно визначається стохастичним ядром [26]:

$$q(D \mid z,a) := \int_{\mathbb{Y}} \mathbf{I}\{H(z,a,y) \in D\}R'(dy \mid z,a), \quad D \in \mathcal{B}(\mathbb{P}(\mathbb{X})),\ z \in \mathbb{P}(\mathbb{X}),\ a \in \mathbb{A}, \tag{6}$$

де $\mathbf{I}$ — індикаторна функція.

Вибір стохастичного ядра H, що задовольняє (5), не впливає на визначення q з (6), оскільки для кожної пари $(z,\ a) \in \mathbb{P}(\mathbb{X}) \times \mathbb{A}$ відображення $\mathbb{H}(z, a, \cdot): \mathbb{Y} \to \mathbb{P}(\mathbb{X})$ визначається $R'(\cdot \mid z, a)$-м.н. однозначно. Аналогічно визначаються стохастичні ядра $R_0, R_0', H_0,\ q_0$ (див. [25]). Тоді $q_0(p)$ є початковим розподілом $z_0 = H_0(p,\ y_0)$, який відповідає розподілу початкового стану p.

Повністю спостережувані МППР визначаються як МППР з параметрами $(\mathbb{P}(\mathbb{X}), \mathbb{A},\ q, \bar{r})$, де $\mathbb{P}(\mathbb{X})$ — простір станів; $\mathbb{A}$ — множина дій, допустимих в усіх станах $z \in \mathbb{P}(\mathbb{X})$; однокрокова функція винагород $r: \mathbb{P}(\mathbb{X}) \times \mathbb{A} \to \overline{\mathbb{R}}$ визначена у такий спосіб:

$$\bar{r}(z,\ a) := \int_{\mathbb{X}} r(x,a)\, z(dx), \quad z \in \mathbb{P}(\mathbb{X}),\ a \in \mathbb{A}; \tag{7}$$

перехідні ймовірності q на $\mathbb{P}(\mathbb{X})$ для заданого $\mathbb{P}(\mathbb{X})\times \mathbb{A}$ визначені в (6).

Для повністю спостережуваних МППР припущення (W^*) можна представити у такому вигляді:

(i) $\bar{r}$ є $\mathbb{K}$-sup-компактною на $\mathbb{P}(\mathbb{X})\times \mathbb{A}$;

(ii) перехідна ймовірність $q(\cdot\,|\,z,a)$ є слабконеперервною на $(z,\ a)\in \mathbb{P}(\mathbb{X})\times \mathbb{A}$.

Позначимо i_t, $t=0,1,\ldots$, історію t-горизонту для повністю спостережуваних МППР, яка називається інформаційним вектором $i_t := (z_0,\ a_0,\ldots, z_{t-1},\ a_{t-1},\ z_t) \in I_t$, $t=0,1,\ldots$, де z_0 — початковий апостеріорний розподіл і $z_t \in \mathbb{P}(\mathbb{X})$ рекурсивно визначається рівнянням (5), $I_t := \mathbb{P}(\mathbb{X})\times (\mathbb{A}\times \mathbb{P}(\mathbb{X}))^t$ для всіх $t=0,1,\ldots$, $I_0 := \mathbb{P}(\mathbb{X})$.

Інформаційна стратегія (I-стратегія) є стратегією повністю спостережуваних МППР, тобто I-стратегія — це така послідовність $\delta = \{\delta_t : t=0,1,\ldots\}$, що $\delta_t(\cdot\,|\,i_t)$ є стохастичним ядром на $\mathbb{A}$ для заданого I_t та $t=0,1$ [22, 23]. Позначимо Δ множину всіх I-статегій. Також розглядатимемо марковську I-стратегію та стаціонарну I-стратегію. Для I-стратегії $\delta = \{\delta_t : t=0,1,\ldots\}$ визначимо стратегію $\pi^\delta = \{\pi_t^\delta : t=0,1,\ldots\}$ в Π:

$$\pi_t^\delta(\cdot\,|\,h_t) := \delta_t(\cdot\,|\,i_t(h_t)) \quad \text{для всіх} \quad h_t \in H_t, \quad t=0,1,\ldots, \tag{8}$$

де $i_t(h_t) \in I_t$ — інформаційний вектор, визначений історією спостережень h_t. Отже, δ і π^δ є еквівалентними в тому сенсі, що π_t^δ має ту ж саму умовну ймовірність на $\mathbb{A}$ для заданої історії спостережень h_t, як і δ_t для історії $i_t(h_t)$.

Якщо δ — оптимальна стратегія для повністю спостережуваних МППР, то π^δ є оптимальною стратегією для частково спостережуваних. Це випливає з того, що $V_{t,\alpha}(p,\ a) = \bar{v}_{t,\alpha}(q_0(p),\ a)$, $t=0,1,\ldots$, і $V_\alpha(p,\ a) = \bar{v}_\alpha(q_0(p),\ a)$. Нехай $z_t(h_t)$ — останній елемент інформаційного вектора $i_t(h_t)$. Застосуємо ті самі позначення для міри, зосередженої в точці. Матимемо, що коли δ є марковською, то з (8) випливає рівність $\pi_t^\delta(h_t) = \delta_t(z_t(h_t))$. До того ж якщо δ є стаціонарною, то $\pi_t^\delta(h_t) = \delta(z_t(h_t))$, $t=0,1,\ldots$ Отже, оптимальна стратегія для повністю спостережуваних МППР визначає оптимальну стратегію для частково спостережуваних МППР. Однак мало знати про умови для частково спостережуваних МППР, за яких існують оптимальні стратегії для відповідних повністю спостережуваних МППР. Зауважимо, що $\bar{v}$ було використане для позначення очікуваних загальних винагород повністю спостережуваних МППР.

Наступна теорема слідує безпосередньо з роботи [18] і застосована до повністю спостережуваних МППР $(\mathbb{P}(\mathbb{X}), \mathbb{A},\ q, -\bar{r})$, оскільки функція ціни стану–дії визначається функцією ціни стану і навпаки. Із цієї теореми випливає, зокрема, алгоритм пошуку оптимальних дій у випадку, коли можна ідентифікувати модель.

Теорема 1. Нехай для повністю спостережуваних МППР $(\mathbb{P}(\mathbb{X}), \mathbb{A},\ q,\ \bar{r})$ задовольняється припущення (W^*) та виконується одне з припущень: (D) або (P). Тоді мають місце такі твердження.

1. Функції $\bar{v}_{t,\alpha}$, $t=0,1,\dots$, і $\bar{v}_{\alpha}$ є напівнеперервними зверху на $\mathbb{P}(\mathbb{X})\times\mathbb{A}$ і $\bar{v}_{t,\alpha}(z)\to\bar{v}_{\alpha}(z)$ у разі $t\to\infty$ для всіх $z\in\mathbb{P}(\mathbb{X})$, $a\in\mathbb{A}$.

2. Для будь-яких $z\in\mathbb{P}(\mathbb{X})$, $a\in\mathbb{A}$ і $t=0,1,\dots$

$$\bar{v}_{t+1,\alpha}(z,\,a)=\bar{\mathrm{r}}\,(z,\,a)+\alpha\int\limits_{\mathbb{P}(\mathbb{X})}\max_{a'\in\mathrm{A}}\bar{v}_{t,\alpha}(z',a')\,q\,(dz'|\,z,a)=$$

$$=\int\limits_{\mathbb{X}}r\,(x,\,a)\,z\,(dx)+$$

$$+\alpha\int\limits_{\mathbb{X}}\int\limits_{\mathbb{X}}\int\limits_{\mathbb{Y}}\max_{a'\in\mathrm{A}}\bar{v}_{t,\alpha}(H(z,a,\,y),a')\,Q\,(dy|\,a,\,x')\,P\,(dx'|\,x,\,a)\,z\,(dx),$$

де $\bar{v}_{0,\alpha}(z,a)=0$ для всіх $z\in\mathbb{P}(\mathbb{X})$, $a\in\mathbb{A}$, і непорожні множини

$$A_{t,\alpha}(z):=\{a\in\mathbb{A}:\bar{v}_{t,\alpha}(z,\,a)=\max_{a'\in\mathbb{A}}v_{t,\alpha}(z,\,a')\},$$

$z\in\mathbb{P}(\mathbb{X})$, $t=0,1,\dots$, задовольняють такі властивості:

а) графік $\mathrm{Gr}\,(A_{t,\alpha})=\{(z,a)\colon z\in\mathbb{P}(\mathbb{X}),\ a\in A_{t,\alpha}(z)\}$, $t=0,1,\dots$, є Борелевою підмножиною $\mathbb{P}(\mathbb{X})\times\mathbb{A}$;

б) якщо $\bar{v}_{t+1,\alpha}(z,\,a)=+\infty$, тоді $A_{t,\alpha}(z,\,a)=\mathbb{A}$, і якщо $\bar{v}_{t+1,\alpha}(z,\,a)<+\infty$, то $A_{t,\alpha}(z,\,a)$ є компактними.

3. Для будь-якого $T=1,2,\dots$ існує оптимальна марковська I-стратегія з T-горизонтом $(\phi_0,\dots,\phi_{T-1})$; якщо для марковської I-стратегії $(\phi_0,\dots,\phi_{T-1})$ з T-горизонтом мають місце включення $\phi_{T-1-t}(z)\in A_{t,\alpha}(z)$, $z\in\mathbb{P}(\mathbb{X})$, $t=0,\dots,T-1$, то ця I-стратегія з T-горизонтом є оптимальною.

4. Для $\alpha\in[0,1)$, якщо виконується припущення (D), і для $\alpha\in[0,1]$, якщо виконується припущення (P),

$$\bar{v}_{\alpha}(z,\,a)=\bar{r}\,(z,\,a)+\alpha\int\limits_{\mathbb{P}(\mathbb{X})}\max_{a'\in\mathrm{A}}\bar{v}_{\alpha}(z',a')\,q\,(dz'|\,z,a)=$$

$$=\int\limits_{\mathbb{X}}r\,(x,\,a)\,z\,(dx)+$$

$$+\alpha\int\limits_{\mathbb{X}}\int\limits_{\mathbb{X}}\int\limits_{\mathbb{Y}}\max_{a'\in\mathrm{A}}\bar{v}_{\alpha}(H(z,a,\,y),a')\,Q\,(dy|\,a,\,x')\,P\,(dx'|\,x,\,a)\,z\,(dx),$$

$z\in\mathbb{P}(\mathbb{X})$, $a\in\mathbb{A}$, і непорожні множини

$$A_{\alpha}(z):=\{a\in\mathbb{A}:\bar{v}_{\alpha}(z,\,a)=\max_{a'\in\mathbb{A}}v_{\alpha}(z,\,a')\},\quad z\in\mathbb{P}(\mathbb{X}),$$

задовольняють такі властивості:

а) графік $\mathrm{Gr}\,(A_{\alpha})=\{(z,a):z\in\mathbb{P}(\mathbb{X}),\ a\in A_{\alpha}(z)\}$ є Борелевою підмножиною $\mathbb{P}(\mathbb{X})\times\mathbb{A}$;

б) якщо $\bar{v}_{\alpha}(z,\,a)=+\infty$, то $A_{\alpha}(z,\,a)=\mathbb{A}$, і якщо $\bar{v}_{\alpha}(z,\,a)<+\infty$, то $A_{\alpha}(z,a)$ є компактними.

5. Для задачі з нескінченним горизонтом існує стаціонарна оптимальна I-стратегія ϕ_{α} та стаціонарна I-стратегія $\phi_{\alpha}^{(*)}$ є оптимальною тоді і тільки тоді, якщо $\phi_{\alpha}^{(*)}(z)\in A_{\alpha}(z)$ для всіх $z\in\mathbb{P}(\mathbb{X})$.

6. Якщо функція $\bar{\mathrm{r}}$ є sup-компактною на $\mathbb{P}(\mathbb{X})\times\mathbb{A}$, то функції $\bar{v}_{t,\alpha}$, $t=1,2,\dots$, і $\bar{v}_{\alpha}$ є sup-компактними на $\mathbb{P}(\mathbb{X})$.

Зауваження 1. Нехай виконується одне з припущень: (D) або (P). Відповідно до [25] припущення (W^{*}) для повністю спостережуваних МППР

$(\mathbb{P}(\mathbb{X}), \mathbb{A}, q, \bar{r})$ в термінах частково спостережуваних МППР $(\mathbb{X}, \mathbb{Y}, \mathbb{A}, P, Q, r)$ забезпечують умови рівномірної напівфеллеровості згортки P та Q і $\mathbb{K}$-sup-компактності на $\mathbb{X} \times \mathbb{A}$ функції однокрокових винагород r. До того ж ці умови є критерійними, тобто вони є інваріантними стосовно переходу від частково спостережуваних МППР $(\mathbb{X}, \mathbb{Y}, \mathbb{A}, P, Q, r)$ до повністю спостережуваних МППР $(\mathbb{P}(\mathbb{X}), \mathbb{A}, q, \bar{r})$ і навпаки.

Достатніми умовами рівномірної напівфеллеровості згортки P та Q є слабка неперервність P та неперервність за варіацією Q або, наприклад, неперервність за варіацією P за умови, що Q не залежить від керувального параметра a.

У разі, коли неможливо ідентифікувати модель або модель є складною, в роботі [31] зазначено, що складність методів динамічного програмування для дисконтованих задач може не бути строго поліноміальною. Тоді алгоритм пошуку наближених оптимальних керувань можна модифікувати, використовуючи методи Q-навчання, «актор–критик», функціональні апроксимації та інші методи глибокого навчання з підкріпленням [7, 27, 28, 32].

СФЕРА МОЖЛИВИХ ЗАСТОСУВАНЬ

Фундаментальні результати щодо розв'язності [3] є відправними позиціями для дослідження послідовних проблем прийняття рішень. Зв'язки з теорією стохастичних ігор [2] і складні задачі дискретної оптимізації [1] особливо актуальні для таких застосувань подвійного призначення як керування радарами/гідролокаторами, логістика і маршрутизація матеріальних засобів. Очікуване застосування отриманих результатів досліджень ґрунтується на забезпеченні кращого розуміння питань керування ризиками під час експлуатації автономних систем.

Сфера потенційних застосувань: управління бортовими радіолокаційними системами раннього попередження [1], багатооглядова класифікація цілей за допомогою гідролокатора із застосуванням підводних протимінних заходів [4], активне керування гідролокатором для відстеження декількох цілей [8] та маршрутизація команд автономних підводних апаратів [10]. Детальний опис застосувань цього класу систем наведено в [11, 29].

ВИСНОВКИ

Для формалізації задачі побудови автономних систем штучного інтелекту визначено простори станів та дій, а також винагороди агента в кожній парі (стан–дія) за умови, що сама модель може бути неідентифікованою або складною.

За допомогою методу послідовних наближень для Q-функцій винагород розроблено математично обґрунтовану методологію побудови наближених за заданою точністю ε-оптимальних стратегій.

Показано, що умови рівномірної напівфеллеровості на згортку ядер переходу та спостережень та $\mathbb{K}$-sup-компактності функції однокрокових винагород для частково спостережуваних МППР є інваріантними стосовно переходу до відповідних повністю спостережуваних МППР.

Отримані результати дають змогу окреслити класи АСШІ (зокрема, подвійного призначення), для яких можна на сучасному рівні математичної строгості обґрунтовано будувати оптимальні та ε-оптимальні стратегії навіть у випадках, коли модель ідентифікується, але обчислювальна складність стандартних алгоритмів динамічного програмування може не бути строго поліноміальною.

СПИСОК ЛІТЕРАТУРИ

1. Feinberg E.A., Bender M.A., Curry M.T., Huang D., Koutsoudis T., Bernstein J.L. Sensor resource management for an airborne early warning radar. *Proceedings of SPIE, Signal and Data Processing of Small Targets*. August 7, Orlando, Florida. 2002. Vol. 4728. P. 145–156.
2. Feinberg E.A., Kasyanov P.O., Zgurovsky M.Z. Continuity of equilibria for twoperson zero-sum games with noncompact action sets and unbounded payoffs. *Annals of Operations Research*. 2022. Vol. 317. P. 537–568. https://doi.org/10.1007/s10479-017-2677-y.
3. Feinberg E.A., Kasyanov P.O., Zgurovsky M.Z. A class of solvable Markov decision models with incomplete information. *Proceedings of the 60th IEEE Conference on Decision and Control*. December 13–15, Austin, Texas. 2021.
4. Myers V., Williams D.P. Adaptive multiview target classification in synthetic aperture sonar images using a partially observable Markov decision process. *IEEE Journal of Oceanic Engineering*. 2012. Vol. 37, N 1. P. 45–55.
5. Piunovskiy A.B. Examples in Markov decision processes. London: Imperial College Press., 2012. 310 p.
6. Puterman M.L. Markov decision processes: Discrete stochastic dynamic programming. London: Wiley, 2005. 684 p.
7. Sutton R.S., Barto A.G. Reinforcement learning: An introduction. Second Edition. London: The MIT Press, 2018. 552 p.
8. Wakayama C.Y., Zabinsky Z.B. Simulation-driven task prioritization using a restless bandit model for active sonar missions. *Proceedings of the 2015 Winter Simulation Conference*. December 6–9, Huntington Beach, California. 2015. Doi: 10.1109/WSC.2015.7408530.
9. Wallis W.A. The statistical research group, 1942–1945. *Journal of the American Statistical Association*. 1980. Vol. 75 (370). P. 320–330.
10. Yordanova V., Griffiths H., Hailes S. Rendezvous planning for multiple autonomous underwater vehicles using a Markov decision process. *IET Radar, Sonar & Navigation*. 2017. Vol. 11, N 12. P. 1762–1769.
11. Silver D., Singh S., Precup D., Sutton R.S. Reward is enough. *Artificial Intelligence*. 2021. Vol. 299. 103535.
12. Kara A.D., Saldi N., Yüksel S. Q-learning for MDPs with general spaces: Convergence and near optimality via quantization under weak continuity. 2021. 25 p. *arXiv preprint* arXiv:2111.06781.
13. Kara A.D., Yüksel S. Convergence of finite memory Q-learning for POMDPs and near optimality of learned policies under filter stability. *Mathematics of Operations Research*. 2022. https://doi.org/10.1287/moor.2022.1331.
14. Parthasarathy K.R. Probability measures on metric spaces. New York: Academic Press, 1967. 288 p.
15. Bertsekas D.P., Shreve S.E. Stochastic optimal control: The discrete-time case. Belmont, MA: Athena Scientific, 1996. 330 p.
16. Hernandez-Lerma O., Lassere J.B. Discrete-time Markov control processes: Basic optimality criteria. New York: Springer, 1996. 216 p.
17. Feinberg E.A., Kasyanov P.O., Zadoianchuk N.V. Berge's theorem for noncompact image sets. *Journal of Mathematical Analysis and Applications*. 2013. Vol. 397, Iss. 1. P. 255–259.
18. Feinberg E.A., Kasyanov P.O., Zadoianchuk N.V. Average-cost Markov decision processes with weakly continuous transition probabilities. *Math. Oper. Res*. 2012. Vol 37, N 4. P. 591–607.

19. Rhenius D. Incomplete information in Markovian decision models. *Ann. Statist*. 1974. Vol. 2, N 6. P. 1327–1334.

20. Yushkevich A.A. Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces. *Theory Probab*. 1976. Vol. 21, N 1. P. 153–158.

21. Dynkin E.B., Yushkevich A.A. Controlled Markov processes. New York: Springer-Verlag, 1979. 292 p.

22. Bertsekas D. Multiagent rollout algorithms and reinforcement learning. arXiv preprint arXiv:1910.00120 (2019). https://doi.org/10.48550/arXiv.1910.00120.

23. Hernandez-Lerma O. Adaptive Markov control processes. New York: Springer-Verlag, 1989. 148 p.

24. Feinberg E.A., Kasyanov P.O., Zgurovsky M.Z. Markov decision processes with incomplete information and semiuniform feller transition probabilities. *SIAM Journal on Control and Optimization*. 2022. Vol. 60, N 4. P. 2488–2513.

25. Sondik E.J . The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Oper. Res*. 1978. Vol. 26, N 2 P. 282–304.

26. Hernandez-Lerma O., Lasserre J.B. Further topics on discrete-time Markov control processes. New York: Springer Science & Business Media, 2012. 276 p.

27. Feinberg, E.A., Kasyanov, P.O., Zgurovsky M.Z. Convergence of value iterations for total-cost mdps and pomdps with general state and action sets. *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*. December 2014. P. 1–8.

28. Szepesvari C. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning.* 2010. Vol. 4 (1). 104 p.

29. Rempel M., Cai J. A review of approximate dynamic programming applications within military operations research. *Operations Research Perspectives*. 2021. Vol. 8. 100204.

30. Department of the Navy. *Science & Technology Strategy for Intelligent Autonomous Systems*. July 2. 2021. 24 p. https://www.nre.navy.mil/media/document/department-navy-science-technology-strategy-intelligent-autonomous-systems.

31. Feinberg E.A., Huang J. The value iteration algorithm is not strongly polynomial for discounted dynamic programming. *Operations Research Letters*. 2014. Vol. 42, N 2. P. 130–131.

32. Arslan G., Yüksel S. Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*. 2017. Vol. 62, N 4. P. 1545–1558.

M.Z. Zgurovsky, P.O. Kasyanov, L.B. Levenchuk

FORMALIZATION OF METHODS FOR THE DEVELOPMENT OF AUTONOMOUS ARTIFICIAL INTELLIGENCE SYSTEMS

Abstract. This paper explores the problem of formalizing the development of autonomous artificial intelligence systems (AAIS), whose mathematical models may be complex or non-identifiable. Using the value-iterations method for Q-functions of rewards, a methodology for constructing of ε-optimal strategies with a given accuracy has been developed. The results allow us to outline classes (including dual-use), for which it is possible to rigorously justify the construction of optimal and ε-optimal strategies even in cases where the models are identifiable but the computational complexity of standard dynamic programming algorithms may not be strictly polynomial.

Keywords: autonomous artificial intelligence systems, Markov decision processes, ε-optimal strategies.

Надійшла до редакції 24.04.2023