**УДК 004.825; 004.891**

**П.О. Кравець**
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

# ЗАОХОЧУВАЛЬНЕ НАВЧАННЯ МУЛЬТИАГЕНТНИХ СИСТЕМ



**Розглянуто проблему заохочувального навчання мультиагентних систем в ігровому формулюванні. Побудовано марківську модель стохастичної гри, сформульовано критерії ігрового навчання, описано Q-метод та відповідний алгоритм розв’язування стохастичної гри, проаналізовано результати комп’ютерної реалізації Q-методу.**

**Ключові слова:** мультиагентна система, стохастична гра, заохочувальне навчання, Q-метод.

**The problem of reinforcement learning of multiagent systems in the game formulation is considered. The Markovian model of stochastic game is constructed, criteria of game learning are formulated, the Q-method and corresponding algorithm of the stochastic game solving are described, results of computer realisation of a Q-method are analysed.**

**Keywords:** multiagent system, stochastic game, reinfocement learning, Q-method.

## Вступ

Функціонування більшості сучасних інформаційних систем (ІС) грунтується на жорстко запрограмованих алгоритмах. За непередбачених впливів зовнішнього середовища у таких системах може порушуватися стабільність режимів роботи, що може призвести до різного роду аварійних ситуацій. Для запобігання критичних станів програмне забезпечення розподілених ІС повинно складатися із взаємодіючих автономних модулів, бути інтелектуальним, пластичним, здатним самостійно відслідковувати зміну станів зовнішнього середовища та приймати своєчасні адекватні рішення. Інакше такі системи повинні будуватися за принципами агентно-орієнтованої методології [1]. Агент ІС – це автономний програмний модуль з елементами штучного інтелекту, здатний самостійно приймати рішення, взаємодіяти з середовищем, іншими агентами та людиною під час розв’язування поставленої задачі. Взаємодія агентів ІС здійснюється у межах комп’ютерної мережі. Популяція агентів комп’ютерної мережі, які розв’язують спільну задачу, називається мультиагентною системою (МАС).

Функціонування МАС [1 – 3], як правило, здійснюється в умовах апріорної невизначеності інформації про стани середовища прийняття рішень та дії інших агентів. У зв’язку з цим стратегії поведінки агентів повинні бути адаптивними [4] за рахунок здатності агентів до самонавчання. Серед методів навчання в умовах невизначеності практичної привабливості набули методи, які грунтуються на заохоченнях [5, 6], оскільки вони не вимагають математичної моделі середовища та забезпечують можливість приймати рішення безпосередньо в процесі навчання. В основу заохочувального навчання покладено механізми рефлекторної поведінки живих організмів з розвинутою нервовою системою. Ефективним методом заохочувального навчання є марківське Q-навчання [7], яке здійснює числову ідентифікацію характеристичної функції динамічної системи у просторі “стан–дія”. Як характеристичну функцію переважно використовують функцію сумарної очікуваної винагороди агента.

Порівняно з одноагентними системами структура, функціонування та дослідження методів багатоагентного Q-навчання значно ускладнюються. За рахунок колективної взаємодії агентів стаціонарне середовище переводиться у клас нестаціонарних. Зміна станів середовища та значення виграшів кожного агента залежать від дій інших агентів. У загальному випадку у МАС агент не може досягти максимального виграшу, який дорівнює його виграшу в одноагентній системі. Оптимальні виграші агентів повинні бути збалансованими і відповідати критеріям вигоди, справедливості, рівноваги. Так, замість критерію скалярної максимізації виграшів одноагентної системи, вводяться

критерії векторної максимізації виграшів МАС, наприклад, рівноваги за Нешем, оптимальності за Парето або ін.

За умови використання методів Q-навчання МАС відбувається ітераційна побудова системи характеристичних Q-функцій у просторі стан-дія, причому приріст елементів цих функцій здійснюється у напрямку досягнення їх колективної рівноваги.

Для побудови МАС необхідно виконати попередні дослідження на основі адекватних математичних моделей, які дають змогу вивчити динаміку системи в умовах невизначеності, побудувати стратегії поведінки агентів, які забезпечують оптимальні техніко-економічні параметри функціонування системи. Враховуючи особливості предметної області, а саме багатоагентність, невизначеність середовища прийняття рішень, антагонізм або конкурентність цілей, комунікативність, координація дій, адаптивність стратегій поведінки агентів, для побудови моделей МАС використаємо математичний апарат теорії стохастичних ігор [8, 9]. Розв'язування стохастичної гри полягає у пошуку таких стратегій агентів, які максимізують їх виграші так, щоб забезпечити певний колективний баланс інтересів усіх гравців. Шукати оптимальні стратегії гравців в умовах невизначеності будемо за методом заохочувального навчання.

Метою роботи є побудова ітераційного методу заохочувального навчання для розв'язування стохастичної гри МАС в умовах невизначеності. Для досягнення мети необхідно розробити модель мультиагентної стохастичної гри, визначити критерії колективної рівноваги, метод та алгоритм розв'язування ігрової задачі.

**Модель мультиагентної стохастичної гри**

Стохастична гра визначається кортежем $(S, p, A_i, r^i \,|\, i \in I)$, $I = \{1, 2, ..., L\}$, де $S = \{s_1, ..., s_M\}$ – множина усіх станів середовища, $p : S \times A \to \Delta(S)$ – функція зміни станів системи, визначена у просторі розподілів імовірностей $\Delta(S)$ на множині $S$, $A_i = \left(a_i(1), ..., a_i(N_i)\right)$ – множина дій або чисті стратегії $i$-го агента; $A = \underset{i \in I}{\times} A_i$ – множина комбінованих дій агентів; $r^i : S \times A \to R$ – функція винагороди $i$-го агента; $I$ – множина агентів; $L$ – число агентів; $M$ – число станів; $N_i$ – кількість стратегій $i$-го агента.

У загальному випадку множини дій $A_i = A_i(s) \;\; \forall i \in I$ та комбінованих дій $A = A(s)$ можуть залежати від станів середовища $s \in S$.

Приймемо марківську модель динаміки станів системи, у якій імовірність зміни станів $p$ залежить тільки від поточного стану середовища і поточних дій агентів:

$$p\left(s_{t+1} = s' \,|\, (s_\tau, b_\tau), \tau = 0, 1, 2, ..., t\right) = p\left(s_{t+1} = s' \,|\, s_t, b_t\right),$$

де $b_t \in A$ – комбінований варіант дії у момент часу $t$.

У кожен момент часу середовище перебуває в одному із станів $s \in S$, і агенти незалежно один від одного вибирають дії $a_i \in A_i$. Після реалізації комбінованого варіанта $a = (a_1, ..., a_L) \in A$ агенти отримують випадкові виграші $r^i$ (інакше – заохочення, підкріплення, стимули), а середовище змінює свій стан згідно із розподілом імовірностей $p(s, a)$ зі значеннями на відрізку $[0, 1]$:

$$\sum_{s' \in S} p(s' \,|\, s, a) = 1.$$

Агент реалізує дії на основі змішаної стратегії

$$\pi_i : S \to A_i,$$

яка визначає імовірності вибору дій $a_i \in A_i$ у кожному стані середовища $\forall s \in S$.

Розподіл $\pi_i \in \Pi_i$ набуває значення на одиничному симплексі $\Pi_i = \left\{ \pi \,\middle|\, \sum_{a_i \in A_i} \pi(s, a_i) = 1, \pi(s, a_i) \geq 0 \right\}$. Якщо $\pi_i(s, a_i) \in \{0, 1\}$, то агент здійснює детермінований вибір варіантів рішень.

Нехай загальний виграш кожного агента визначається функцією дисконтованих сумарних виграшів:

$$\Upsilon_i = \sum_{t=0}^{\infty} \gamma^t r_t^i \ , \tag{1}$$

де $\gamma \in (0, 1]$ – параметр дисконтування.

Мета $i$-го агента полягає у максимізації функції (1) за рахунок формування ефективної стратегії $\pi^i$:

$$V_\pi^i(s) = E_\pi \left[ \Upsilon_i \mid s_0 = s \right] \to \max_{\pi_i}, \ \forall i \in I \ , \tag{2}$$

де $\pi = (\pi_1, ..., \pi_L)$; $E$ – символ математичного сподівання.

Розв'язування стохастичної гри полягає у визначенні стратегій поведінки агентів $\pi_i^*$ ($\forall i \in I$), які забезпечують виконання однієї з умов колективної оптимальності, наприклад:
1) рівноваги за Нешем:

$$V^i(s, \pi_1^*, \pi_2^*, ..., \pi_L^*) \geq V^i(s, \pi_1^*, \pi_2^*, ..., \pi_{i-1}^*, \pi_i \ , \pi_{i+1}^*, ..., \pi_L^*);$$

2) оптимальності за Парето:

$$V^i(s, \pi^*) \geq V^i(s, \pi) \, .$$

**Навчання стохастичної гри**

Обчислення $V_\pi^i(s)$ може бути виконано у рекурсивній формі, відомій у літературі як рівняння Беллмана. Враховуючи (1), після нескладних перетворень отримаємо:

$$V_\pi(s \mid s_t = s) = E(r_t) + \gamma \sum_{k=0}^{\infty} \gamma^k E(r_{t+k+1}) = E(r_t) + \gamma V_\pi(s_{t+1}) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi(s)) V_\pi(s') \, . \tag{3}$$

де $s'$ – можливі майбутні стани системи.

Метою агента є знаходження функції вибору стратегій $\pi^*$, яка максимізує функцію (2) для всіх станів середовища:

$$\forall \pi \, \forall s \in S \qquad V^{\pi^*}(s) \geq V^\pi(s) \, .$$

Оскільки вибирають варіанти дій випадково, то для порівняння ефективності дій, коли система перебуває у стані $s \in S$, поточні виграші корисно отримати з (2). Для цього використовується спеціально побудована $Q$-функція середніх виграшів, яка визначає ціну дії – сумарний виграш агента, який у стані $s$ вибрав дію $a$:

$$Q_\pi(s, a) = E_\pi \left[ R \middle| s_0 = s, a_0 = a \right] \, . \tag{4}$$

Тут $Q_\pi(s, a)$ є табличною функцією значень варіантів дій $a$ у станах $s$.

Аналогічно до (3) отримаємо:

$$Q_\pi(s, a) = r(s, a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V_\pi(s') \, . \tag{5}$$

Дотримання принципу оптимальності Беллмана забезпечує оптимальний виграш агента з досягнутого поточного стану $s \in S$ в усі майбутні моменти часу. Застосування цього принципу для усіх станів забезпечує отримання глобального оптимального розв'язку.

Для оптимальної функції вибору стратегій $\pi^*$ для кожного стану $s \in S$ отримаємо:

$$V_{\pi^*}(s) = \max_{a \in A} \left[ r(s, a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V_{\pi^*}(s') \right]. \tag{6}$$

З (6) можна отримати оптимальну функцію вибору стратегій

$$\pi^*(s) = \arg\max_{a \in A} Q_{\pi^*}(s, a) \, . \tag{7}$$

Оптимізацію (7) можна виконати методами динамічного програмування [10].

За аналогією до одноагентного навчання [5 – 7] визначимо матрицю виграшів $i$-го гравця з урахуванням поточного та майбутніх виграшів у напрямку руху до оптимального колективного стану у просторі $S \times A$:

$$Q_*^i(s, a_1, ..., a_L) = r^i(s, a_1, ..., a_L) + \gamma \sum_{s' \in S} p(s' \mid s, a_1, ..., a_L) \cdot V^i(s', \pi_1^*, ..., \pi_L^*),$$

де $Q_*^i(s, a_1, ..., a_L)$ – загальний дисконтований виграш $i$-го гравця за умови вибору гравцями дій $(a_1, ..., a_L)$ у стані $s$ згідно з оптимальною стратегією гри $\pi^* = (\pi_1^*, ..., \pi_L^*)$.

В умовах апріорної невизначеності імовірностей переходів між станами системи $p(s, a_1, ..., a_L)$ та функції виграшів $r^i(s, a_1, ..., a_L)$ для обчислення елементів матриць виграшів використовують метод ітераційного $Q$-навчання [9]:

$$Q_{t+1}^i(s, a_1, ..., a_L) = (1 - \alpha_t) Q_t^i(s, a_1, ..., a_L) + \alpha_t [r_t^i + \gamma V_t^i(s_{t+1})], \tag{8}$$

де $\alpha_t \in (0, 1)$ – параметр навчання; $V_t^i(s_{t+1})$ – оператор вартості стану системи у напрямку оптимального колективного розв'язку.

Вигляд оператора $V_t^i(s_{t+1})$ визначається умовою колективної рівноваги, наприклад:

$V_t^i(s_{t+1}) = MM\left(Q_t(s_{t+1})\right)$ – максимінна рівновага;

$V_t^i(s_{t+1}) = NE\left(Q_t^i(s_{t+1})\right)$ – рівновага за Нешем;

$V_t^i(s_{t+1}) = BR\left(Q_t^i(s_{t+1})\right)$ – найкраща відповідь агента;

$V_t^i(s_{t+1}) = CE\left(Q_t^i(s_{t+1})\right)$ – корельована рівновага;

$V_t^i(s_{t+1}) = PE\left(Q_t^i(s_{t+1})\right)$ – оптимальність за Парето.

Наведений перелік можна доповнити іншими вже відомими та новими станами рівноваги, які визначатимуть цільовий аспект функціонування розподіленої динамічної системи.

Метод (8) можна застосувати для розв'язування гри одного агента з природою як часткового випадку $N$-агентної стохастичної гри, якщо $I = \{i\}, \mid I \mid = 1, A = A_i$, коли

$$V_t^i(s_{t+1}) = \max_{b \in A_i} Q_t^i(s', b),$$

де $s' = s_{t+1}$.

Максимінна рівновага (MM, Maximin Equilibrium) існує у грі двох агентів з нульовою сумою функцій їхніх виграшів:

$$V^1(s_{t+1}) = \max_{\pi_1 \in \Pi_1} \min_{a_2 \in A_2} \sum_{a_1 \in A_1} \pi_1(s', a_1) Q_t(s', a_1, a_2) = -V^2(s_{t+1}) .$$

Рівновага за Нешем (NE, Nash Equilibrium) визначається незалежним розподілом стратегій гравців, які вибирають власні стратегії самостійно, незважаючи на вибір інших агентів. У ситуації рівноваги за Нешем у змішаних стратегіях $\pi^{NE}(s) = \left(\pi_1^{NE}(s), ..., \pi_L^{NE}(s)\right)$ кожному агенту не вигідно відхилитися від власної оптимальної стратегії $\pi_i^{NE}(s)$, якщо інші агенти дотримуються точки рівноваги [9]:

$$\sum_{a \in A} Q_t^i(s', a) \pi_i^{NE}(a_i) \prod_{j \neq i}^{L} \pi_j^{NE}(s', a_j) \geq \sum_{a \in A} Q_t^i(s', a) \tilde{\pi}_i(a_i) \prod_{j \neq i}^{L} \pi_j^{NE}(s', a_j), \tag{9}$$

де $a = (a_1, ..., a_L)$; $\pi_i^{NE}, \tilde{\pi}^i \in \Pi_i$.

Метод (8) забезпечує виконання умови (9) коли поточне значення оператора вартості станів системи визначається у точці $\pi^{NE}$ рівноваги за Нешем:

$$NE\left(Q_t^i(s_{t+1})\right) = \sum_{a\in A} Q_t^i(s',a)\prod_{j=1}^{L}\pi_j^{NE}(s',a_j).$$

Множина точок NE-рівноваги у змішаних стратегіях є опуклим компактом і може бути обчислена за допомогою методів лінійного програмування (для біматричних ігор) або на основі розв'язування системи полілінійних рівнянь, що визначають умову доповняльної нежорсткості:

$$\sum_{a_{-i}\in A_{-i}} Q_t^i(s,a_{-i},a_i)\prod_{j\neq i}^{L}\pi_j(s,a_j) = \sum_{a\in A} Q_t^i(s,a)\prod_{j=1}^{L}\pi_j(s,a_j),$$

$$\forall i\in I,\ \forall a_i\in A_i,\ \forall s\in S,\ \pi_i(s,a_i)>0,$$

$$\sum_{a_i\in A_i}\pi_i(s,a_i)=1.$$

На відміну від рівноваги за Нешем, метод найкращої відповіді (BR, Best Response) формує оптимальну стратегію агента у відповідь на дії усіх інших агентів. Відповідний оператор вартості стану системи у методі (8) має вигляд [11]:

$$BR\left(Q_t^i(s_{t+1})\right) = \max_{\pi_i}\left(\sum_{a\in A} Q_t^i(s',a)\prod_{j=1}^{L}\pi_j(s',a_j)\right).$$

Корельована рівновага (CE, Correlated Equilibrium) узагальнює рівновагу за Нешем, допускаючи залежність стратегій гравців. Для цього у системі колективного прийняття рішень присутній арбітр, який згідно з узагальненим розподілом $\sigma\in\Delta(A)$ ($\sum_{a\in A}\sigma(a)=1$, $A=\underset{i=1}{\overset{L}{\times}}A_i$) рекомендує гравцям обрати для реалізації дії, які утворюють комбінований варіант $a=(a_1,...,a_L)$. Гравець з номером $i$ отримує інформацію тільки про компоненту $a_i$ комбінованого варіанта $a\in A$. Цей сигнал сприймається $i$-м гравцем як необов'язкова пропозиція виконати дію $a_i$. Кожен гравець таємно і незалежно вибирає у момент часу $t$ варіант дії $a_i$, можливо, відмінний від запропонованого варіанта, і отримує поточний виграш $r^i(s_t,a_t)$, який є функцією поточного стану системи $s_t$ та комбінованого варіанта $a_t\in A$. Далі середовище переходить у новий стан $s_{t+1}$ згідно із розподілом імовірностей $p(s_{t+1}\mid s_t,a_t)$, і процес повторюється у момент часу $t+1$.

Корельована рівновага визначається об'єднаним розподілом стратегій гравців $\sigma\in\Delta(A)$, коли кожен агент не має мотивацій, щоб відхилитися від домовленостей в односторонньому порядку [12]:

$$\sum_{a_{-i}\in A_{-i}}\sigma^{CE}(a_{-i}\mid a_i)Q^i[s',(a_{-i},a_i)] \geq \sum_{a_{-i}\in A_{-i}}\sigma^{CE}(a_{-i}\mid a_i)Q^i[s',(a_{-i},\tilde{a}_i)],$$

де $A_{-i}=\underset{j=1,\,j\neq i}{\overset{L}{\times}}A_j$, $A=A_{-i}\times A_i$, $a_{-i}\in A_{-i}$, $a=(a_{-i},a_i)\in A$, $a_i,\tilde{a}_i\in A_i$, $\sigma^{CE}(a_i)=\sum_{a_{-i}\in A_{-i}}\sigma^{CE}(a_{-i},a_i)$,

$\sigma^{CE}(a_{-i}\mid a_i)=\sigma^{CE}(a_{-i},a_i)\big/\sigma^{CE}(a_i)$, $\sigma^{CE}(a_i)>0$.

Усі точки рівноваги за Нешем є точками корельованої рівноваги. Якщо $\forall i\in I$ $\sigma^{CE}(a_{-i}\mid a_i)=\sigma^{CE}(a_{-i}\mid\tilde{a}_i)$, $\forall a_i,\tilde{a}_i\in A_i$, $\forall a_{-i}\in A_{-i}$, $\forall\sigma^{CE}(a_i),\sigma^{CE}(\tilde{a}_i)>0$, то корельована рівновага є також рівновагою за Нешем.

Для розв'язування гри методом (8) оператор вартості $V_t^i(s_{t+1})$ стану системи визначається у точці $\sigma^{CE}$ корельованої рівноваги:

$$CE(Q_t^i(s_{t+1})) = \sum_{a\in A}\sigma^{CE}(a)Q_t^i(s',a).$$

Множина точок CE-рівноваги є непорожньою, опуклою та компактною і може бути ефективно обчислена за допомогою методів лінійного програмування. У випадку максимізації сумарного виграшу гравців задача лінійного програмування для знаходження $\sigma^{CE}$ може бути сформульована так:

$$\sum_{a\in A}\sigma(a)\sum_{i=1}^{L}Q^i(s,a)\to\max_{\sigma},$$

$$\sum_{a_{-i}\in A_{-i}} \sigma(a_{-i},a_i)\Big(Q^i[s',(a_{-i},a_i)]-Q^i[s',(a_{-i},\tilde{a}_i)\Big)\ge 0,$$

$$\forall i\in I, \forall a_i\in A_i, \forall \tilde{a}_i\in A_i, \forall s\in S,$$

$$\sigma(a)>0 \ \forall a\in A,$$

$$\sum_{a\in A}\sigma(a)=1.$$

На основі розподілу $\sigma(a) \ \ \forall a\in A$ визначаються власні стратегії гравців $\pi_i \ \forall i\in I$. Можливі різні варіанти переходу від $\sigma$ до $\pi_i$ залежно від вигляду стратегій та ступеня інформованості гравців. Так, чисті стратегії визначають максимальне значення оператора $CE(Q^i(s))$:

$$a_i=\arg\max_{\sigma} CE(Q^i(s)).$$

Враховуючи, що $\sum_{a_i\in A_i}\sigma(a_i)=1$, можна прийняти, що змішані стратегії $\pi_i(a_i)=\sigma(a_i)=\sum_{a_{-i}\in A_{-i}}\sigma(a_{-i},a_i)$, або:

$$\pi_i(a_i)=\sum_{a_{-i}\in A_{-i}}\sigma(a_{-i})Q^i(s,a_{-i},a_i)\Big/CE(Q^i(s)).$$

Рівновага (оптимальність) за Парето (PE, Pareto Equilibrium) існує у грі зі спільними інтересами (Common-Interest Markov Game), коли матриці виграшів є однаковими для усіх гравців $Q_t^i(s,a)=Q_t^j(s,a)\forall i,j\in I$, $\forall s\in S, \forall a\in A$ [12].

Гру з різними матрицями виграшів можна перетворити на гру зі спільними інтересами за допомогою згортки

$$PE\Big(Q_t(s_{t+1})\Big)=\sum_{k=1}^{L}\lambda_k\sum_{a\in A}Q_t^k(s',a)\prod_{j=1}^{L}\pi_j^{PE}(s',a_j),$$

де $\lambda_j>0$ $(j=1..L)$.

Шукають PE-розв'язок гри незалежним вибором стратегій агентів, аналогічно до пошуку NE-розв'язку.

Багатоагентна гра є оптимальною за Парето, якщо не існує спільної стратегії гравців, яка дає змогу покращити виграші усіх гравців:

$$Q_t^i(s,\pi^{PE})\ge Q_t^i(s,\pi).$$

Парето-оптимальні змішані стратегії $\pi^{PE}(s)=\Big(\pi_1^{PE}(s),...,\pi_L^{PE}(s)\Big)$ можна отримати максимізацією згортки увігнутих (вгору) функцій виграшів:

$$\sum_{k=1}^{L}\lambda_k\sum_{a\in A}Q_t^k(s',a)\prod_{j=1}^{L}\pi_j(s',a_j)\to\max_{\pi}.$$

Для обчислення оптимальних колективних стратегій $\pi^*(s)=\Big(\pi_1^*(s),...,\pi_L^*(s)\Big)$ (NE, CE, PE або ін.) агенту з номером $i$ необхідно знати $Q$-функції усіх агентів: $Q(s)=\Big(Q_t^1(s),...,Q_t^L(s)\Big)$. За відсутності такої інформації кожен агент повинен оцінити значення $Q$-функцій у процесі навчання. Для цього $i$-й агент спостерігає поточні виграші інших агентів і модифікує оцінки їх $Q$-функцій згідно з (8).

Для забезпечення збіжності методу (8) до однієї із точок колективної рівноваги необхідно накласти обмеження на швидкість зміни його регульованих параметрів. Загальні обмеження є такими [7]:

$$\sum_{t=0}^{\infty}\alpha_t=\infty, \qquad \sum_{t=0}^{\infty}\alpha_t^2<\infty, \tag{10}$$

де $\alpha_t=t^{-\kappa}$ $(\kappa>0)$ – монотонно спадні невід'ємні послідовності дійсних величин.

**Алгоритм розв'язування стохастичної гри**

1. Задати початковий момент часу $t=0$; початкові значення матриць виграшів $Q_t^i(s,a_1,...,a_L)=\varepsilon \quad \forall s\in S,\ \forall a_i\in A_i,\ \forall i\in I$, де $0<\varepsilon<<1$ – мале додатне значення; значення параметра дисконтування виграшів $\gamma\in(0,1]$; початковий стан системи $s_0$.

2. Виконати випадковий вибір дій агентів $a=(a_1,...,a_L)$ на основі стратегій $\pi=(\pi_1,...,\pi_L)$. Значення стратегій можна обчислити на основі поточних оцінок матриць виграшів $\forall i\in I$:

$$\pi_i(s,a_i)=\sum_{a_{-i}\in A_{-i}} Q_t^i(s,a_{-i},a_i)\Big/\sum_{a\in A} Q_t^i(s,a),\ \forall a_i\in A_i.$$

3. Отримати поточні виграші агентів $r_t=(r_t^1,...,r_t^L)$.

4. Визначити новий стан системи $s_{t+1}=s_t(a_1,...,a_L)$.

5. Обчислити функцію $V_t^i(s_{t+1})$ згідно із конкретною умовою колективної рівноваги (NE, CE, PE або ін.).

6. Модифікувати матриці виграшів $Q_{t+1}=\left(Q_{t+1}^i(s_t,a_t)\,|\,i=1..L\right)$ згідно з (8).

7. Якщо $\left\|Q_{t+1}^i-Q_t^i\right\|<\varepsilon\ \forall i=1..L$, то задати $t:=t+1$ і перейти на крок 2.

8. Вивести розраховані значення матриць виграшів $Q(s)=\left(Q^1(s),...,Q^L(s)\right)$ та стратегій $\pi(s)=(\pi_1(s),...,\pi_L(s))\ \ \forall s\in S$. Кінець.

**Результати комп'ютерного моделювання**

Виконаємо розв'язування стохастичної гри двох агентів з двома чистими стратегіями у середовищі з двома станами. Матриці середніх виграшів такої гри подано у таблиці.

**Матриці виграшів гравців**

| Стани | Стратегії | Перший гравець | | Другий гравець | |
|---|---|---|---|---|---|
| $s_1$ | × | $\pi_2(s_1,a_2[1])$ | $\pi_2(s_1,a_2[2])$ | $\pi_2(s_1,a_2[1])$ | $\pi_2(s_1,a_2[2])$ |
| | $\pi_1(s_1,a_1[1])$ | 0.4 | 0.1 | 0.9 | 0.2 |
| | $\pi_1(s_1,a_1[2])$ | 0.1 | 0.9 | 0.2 | 0.9 |
| $s_2$ | × | $\pi_2(s_2,a_2[1])$ | $\pi_2(s_2,a_2[2])$ | $\pi_2(s_2,a_2[1])$ | $\pi_2(s_2,a_2[2])$ |
| | $\pi_1(s_2,a_1[1])$ | 0.4 | 0.6 | 0.5 | 0.2 |
| | $\pi_1(s_2,a_1[2])$ | 0.6 | 0.8 | 0.6 | 0.7 |

В умовах невизначеності елементи матриць середніх виграшів $\left[v^i(s,a)\right]_{\substack{\forall s\in S\\ \forall a\in A}}$ апріорі невідомі і доступні для спостереження у вигляді випадкових поточних значень

$$r^i(s,a)=Normal(v^i(s,a),d^i(s,a)),$$

розподілених за нормальним законом законом з математичним сподіванням $v^i(s,a)$ та дисперсією $d^i(s,a)$.

Нормально-розподілені випадкові величини отримано за допомогою суми дванадцяти рівномірно-розподілених випадкових чисел $\omega\in[0,1]$:

$$r^i(s,a)=v^i(s,a)+\sqrt{d^i(s,a)}\left(\sum_{j=1}^{12}\omega_j-6\right), \qquad (11)$$

де $d^i(s,a)=d>0\ \forall s\in S,\forall a\in A$.

Якщо у момент часу $t$ система перебувала у стані $s\in S$, то після реалізації чистих стратегій $a=(a_1,...,a_L)$, де $a\in A=\underset{i=1}{\overset{L}{\times}}A^i$, агенти отримують поточні виграші $r_t^i(s,a)$, обчислені згідно з (11).

Після отримання поточних виграшів кожен агент перераховує відповідний елемент $Q$ – матриці згідно з модифікованим для умов невизначеності алгоритмом $BR$:

$$Q^i_{t+1}(s,a_1,...,a_L) = (1-\alpha_t)Q^i_t(s,a_1,...,a_L) + \alpha_t(r^i_t + \gamma \max_{a_i} Q^i_t(s,a_1,...,a_L)). \quad (12)$$

На основі $Q$-матриць обчислюються поточні значення змішаних стратегій за методом Больцмана:

$$\pi_i(a_i(k)\,|\,s) = e^{Q^*_i(s,a_i(k))/T} \Big/ \sum_{j=1}^{N_i} e^{Q^*_i(s,a_i(j))/T}, \; k = 1..N_i, \quad (13)$$

де $Q^*_i(s,a_i(k)) = \max_{a_{-i}} r^i(a_{-i},a_i(k))$, $a_{-i} \in A_{-i}$, $A_{-i} = \underset{\substack{j=1 \\ j\neq i}}{\overset{L}{\times}} A^j$, $T > 0$ – температурний коефіцієнт.

Елементи вектора змішаних стратегій $\pi_i$ задають дискретний розподіл, за яким визначаються значення випадкових чистих стратегій $i$-го агента у наступний момент часу:

$$a_i(s) = \left\{ A^i(s,k) \,\middle|\, k = \arg\left( \min_k \sum_{j=1}^{k} \pi_i(s,a_i(j)) > \omega \right), k = 1..N_i \right\} \; \forall s \in S, \; \forall i \in I, \quad (14)$$

де $\omega \in [0,1]$ – випадкова величина з рівномірним розподілом.

Зміна станів динамічної системи визначається дискретним розподілом $p(s'\,|\,s,a) = p$ $\forall s \in S, \forall a \in A$:

$$s = \left\{ S(k) \,\middle|\, k = \arg\left( \min_k \sum_{j=1}^{k} p(j) > \omega \right), k = 1..M \right\}. \quad (15)$$

Нехай стани $s \in S$ системи змінюються з однаковими імовірностями $p(s'\,|\,s,a) = (|S|_k^{-1}, k = 1..M)$ $\forall s \in S, \forall a \in A$, тобто $p(s'\,|\,s,a) = (0.5;\, 0.5)$ для $|S| = 2$.

Середні виграші агентів обчислюються із врахуванням імовірностей переходу середовища із одного стану в інший:

$$V^i = \sum_{s \in S} p(s) V^i(s), \; i = 1..L,$$

де $V^i(s) = \sum_{a \in A} v^i(s,a) \prod_{j=1}^{L} \pi_j(s,a)$ – середній виграш агента у стані $s \in S$.

Траєкторії зміни стратегій агентів у межах одиничного симплексу та вигляд функцій середніх виграшів $V^i(s)$, які відповідають даним таблиці, зображено на рис. 1 та 2.

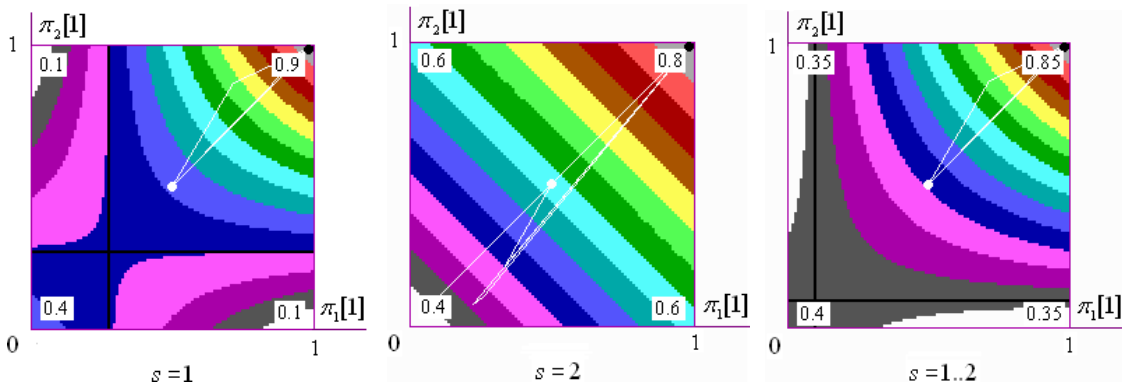


*Рис. 1. Функції середніх виграшів першого агента*

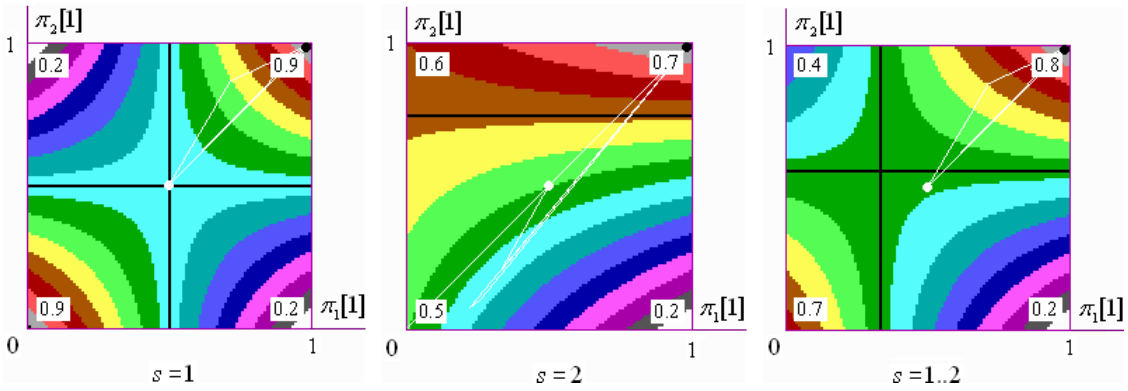


*Рис. 2. Функції середніх виграшів другого агента*

Метод BR (12 – 15) забезпечує розв'язування стохастичної гри на вершинах одиничного симплекса з максимальним значенням функції середніх виграшів. Відсоток варіантів досягнення оптимального розв'язку гри залежить від абсолютної різниці між двома найбільшими послідовними значеннями функцій середніх виграшів.

Збіжність методу оцінюється похибкою виконання умови доповняльної нежорсткості [13], зваженої змішаними стратегіями:

$$\Delta = L^{-1}\sum_{i\in I}\left\|\pi_i - \tilde{\pi}_i\right\|^2 ,$$

де $\tilde{\pi}_i = diag(\pi_i)\nabla V^i / V^i$; $diag(\pi_i)$ – діагональна квадратна матриця порядку $N_i$, сформована з елементів вектора $\pi_i$; $\nabla V^i = \left(V^i[j]\middle| j = 1..N_i\right)$ – векторна функція середніх виграшів для фіксованих чистих стратегій $i$-го гравця; $V^i = \sum_{j=1}^{N_i} V^i[j]\pi_i[j]$ – функція середніх виграшів $i$-го гравця; $\|*\|$ – евклідова норма вектора.

Умова доповняльної нежорсткості характеризує розв'язки гри у змішаних стратегіях за Нешем. Зважувана умова додатково враховує розв'язки гри у чистих стратегіях.

Графіки функцій середніх виграшів $\Upsilon$ та норми відхилення змішаних стратегій від їх цільових значень $\Delta$ подано на рис. 3. Виграші усереднено за кількістю гравців:

$$\Upsilon = L^{-1}\sum_{i=1}^{L}\Upsilon_i ,$$

де $\Upsilon_i \geq 0$ – дисконтовані поточні виграші $i$-го гравця.

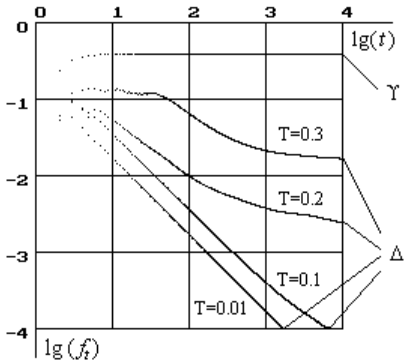


*Рис. 3. Характеристики збіжності ігрового $Q$-методу*

Спадання графіка норми відхилення $\Delta$ змішаних стратегій від їх цільового значення свідчить про збіжність ігрового $Q$-методу.

Значення температурного коефіцієнта $T$ значно впливає на збіжність ігрового методу. Швидкість збіжності визначається стрімкістю спадання графіка функції $\Delta$, яку можна оцінити величиною гострого кута лінійної апроксимації графіка функції $\Delta$ з віссю часу. Зі зростанням $T$ швидкість збіжності ігрового $Q$-методу зменшується.

**Висновки**

Розглянутий метод заохочувального навчання (8) у детермінованому варіанті вимагає знання кожним агентом $Q$-функцій усіх інших агентів. Ці функції використовуються агентами для визначення стратегій, які забезпечують динаміку методу у напрямку точок колективної рівноваги.

Значення $Q$-функцій можна отримати в результаті обміну інформацією між агентами. Якщо інтегрована інформація про $Q$-функції не доступна агенту, то він повинен визначити їх значення самостійно в процесі навчання, спостерігаючи за поточними виграшами інших агентів і виконуючи оцінювання $Q$-функцій згідно із (8). Якщо такі спостереження не можливі, то агенти можуть виконати рефлексивні оцінки $Q$-функцій інших агентів [14].

Інший спосіб побудови алгоритмів заохочувального навчання агентів в умовах невизначеності полягає у застосуванні методу стохастичної апроксимації [15] для відповідної умови колективної рівноваги.

Практичне використання методів ігрового заохочувального навчання вимагає їх попереднього аналізу для визначення умов збіжності до стану колективної рівноваги. Такі дослідження здійснюються на основі оцінювання послідовностей випадкових величин [4, 16], які характеризують поточні відхилення стратегій гравців від їхніх оптимальних значень.

Швидкість збіжності ігрового $Q$-методу заохочувального навчання визначається параметрами $\alpha_t$ та $T$. Параметр $\alpha_t$ повинен задовольняти загальні умови стохастичної апроксимації (10). Величина параметра $T$ залежить від абсолютних значень елементів $Q$-матриць. Експериментально встановлено, що для заданих у роботі матриць середніх виграшів збіжність ігрового $Q$-методу забезпечується при $T \in (0, 0.2]$ у діапазоні значень параметра $\alpha_t = t^{-\kappa}$, $\kappa \in (0, 1]$. Найбільша швидкість збіжності ігрового методу заохочувального навчання досягається при $T = 10^{-2}$.

*1. Тарасов В.Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика / В.Б. Тарасов. – М.: Эдиториал УРСС, 2002. – 352 с. 2. Wooldridge M. An Introduction to Multiagent Systems / M. Wooldridge. – John Wiley & Sons, 2002. – 366 pp. 3. Weiss, G. Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence / G. Weiss, editor. – Springer Verlag, Berlin, 1996. – 643 pp. 4. Назин А.В. Адаптивный выбор вариантов: Рекуррентные алгоритмы / А.В. Назин, А.С. Позняк. – М.: Наука, 1986. – 288 с. 5. Kaelbling, Leslie. Reinforcement learning: A survey / Leslie Kaelbling, Michael L. Littman, Andrew W. Moore. Journal of Artificial Intelligence Research. – 1996. – No. 4. – PP. 237–285. 6. Sutton, R. S. Reinforcement Learning: An Introduction / Richard S. Sutton, Andrew G. Barto. – MIT Press, 1998. – 322 pp. 7. Watkins, C.J.C.H. Q-Learning / C.J.C.H. Watkins, P. Dayan // Machine Learning. – Kluwer Academic Publishers, Boston. – 1992. – No. 8. – PP. 279–292. 8. Fudenberg, D. The Theory of Learning in Games / D. Fudenberg, D.K. Levine. – Cambridge, MA: MIT Press, 1998. – 292 pp. 9. Hu, J. Nash Q-learning for general-sum stochastic games / J. Hu, M. P. Wellman // Machine Learning Research. – 2003. – No. 4. – PP. 1039 – 1069. 10. Puterman, M. L. Markov Decision Processes: Discrete Stochastic Dynamic Programming / M. L. Puterman. – John Wiley & Sons, New York, 2005. – 649 pp. 11. Weinberg, M. Best-Response Multiagent Learning in Non-Stationary Environments / Michael Weinberg, Jeffrey S. Rosenschein // AAMAS'04. – New York, USA. – July 19 – 23, 2004. 12. Greenwald, A. Correlated Q-learning / A. Greenwald, K. Hall // Proceedings of the Twentieth International Conference on Machine Learning. – 2003. – PP. 242–249. 13. Мулен Э. Теория игр с примерами из математической экономики / Э. Мулен. – М.: Мир, 1985. – 200 с. 14. Новиков, Д.А. Рефлексивные игры / Д.А. Новиков, А.Г. Чхартишвили. – М.: СИНТЕГ, 2003. – 149 с. 15. Вазан, М. Стохастическая аппроксимация / М. Вазан. – М.: Мир, 1972. – 295 с. 16. Невельсон, М.Б. Стохастическая оптимизация и рекуррентное оценивание / Невельсон М.Б., Хасьминский Р.З. – М.: Наука, 1972. – 304 с.*