УДК 004.852; 004.94

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

МАТРИЧНА СТОХАСТИЧНА ГРА З Q-НАВЧАННЯМ

Розроблена модель матричної стохастичної гри для прийняття рішень в умовах невизначеності. Запропоновано метод Q-навчання для розв’язування стохастичної гри з апріорі невідомими матрицями виграшів. Виконано формулювання ігрової задачі, описано марківський рекурентний метод та алгоритм для її розв’язування. Отримано та проаналізовано результати комп’ютерного моделювання стохастичної гри з Q-навчанням.

Ключові слова: стохастична гра, умови невизначеності, Q-навчання, марківський рекурентний метод.

The model of matrix stochastic game for decision-making in the conditions of uncertainty is developed. The method of Q-learning for stochastic game solving with a priori unknown gains matrices is offered. The formulation of a game problem is executed. The Markovian recurrent method and algorithm for the game solving are described. Results of computer modelling of stochastic game with Q-learning are received and analysed.

Key words: stochastic game, uncertainty conditions, Q-learning, Markovian recurrent method.

Вступ. Загальна постановка проблеми

Стохастичні ігрові моделі використовуються для розв’язування задач, пов’язаних із необхідністю прийняття рішень в умовах невизначеності – в біології, психології, соціології, політичній науці, військовій справі, економіці, маркетингу, екології, інформаційних, програмних та технічних системах [1, 2]. Характерними особливостями таких задач є:

1) розподіленість або багатопараметричність середовища прийняття рішень;

2) внутрішня стохастичність середовища;

3) повна або часткова відсутність апріорної інформації про середовище прийняття рішень;

4) керованість середовища та можливість розподіленої реалізації варіантів керування;

5) визначеність векторної мети керування або прийняття рішень;

6) дискретність та скінченність множини варіантів прийняття рішень;

7) стохастична незалежність вибору варіантів рішень у просторі та у часі;

8) можливість багатократного повторення реалізацій варіантів дій гравців на безмежному відрізку часу;

9) розподілений локально-залежний характер формування та збору інформації для статистичної ідентифікації середовища прийняття рішень;

10) можливість застосування розподіленого ігрового алгоритму, який забезпечує досягнення області компромісних рішень;

11) реалізація ігрового алгоритму в реальному масштабі часу;

12) можливість визначення моментів зупинки ігрового алгоритму для його практичного застосування.

Матрична стохастична гра задається множиною гравців, структурами їх локальних взаємодій, матрицями розподілів випадкових виграшів або програшів, множинами чистих та змішаних стратегій, правилами прийняття рішень. Чисті стратегії визначають множини варіантів рішень, а змішані стратегії – умовні імовірності вибору чистих стратегій. Гравці, наділені здатністю автономного вибору варіантів рішень, називаються агентами прийняття рішень [3, 4].

Нижче, для однозначності, будемо розглядати стохастичну гру з максимізацією функцій виграшів. На відміну від детермінованої гри, в умовах невизначеності гравцям апріорі не відомі їх матриці виграшів. Учасники стохастичної гри отримують тільки поточні реакції середовища у відповідь на реалізацію їх чистих стратегій. Чисті стратегії гравців визначаються випадково на основі імовірнісних розподілів, що задаються змішаними стратегіями.

Для знаходження розв'язків стохастичної гри в умовах невизначеності використовуються ітераційні методи. Повторення кроків гри необхідне для збору інформації про ефективність стратегій гравців у ході оптимізації їх цільових функцій.

Відомі рекурентні методи розв'язування стохастичної гри ґрунтуються на пошуку оптимальних значень змішаних стратегій у межах одиничних симплексів [5]. Належність змішаних стратегій одиничному симплексу забезпечується проекційним оператором. Такі методи є простими для програмування, не вимагають обміну інформації між гравцями і в умовах невизначеності забезпечують степеневий порядок швидкості збіжності.

Крім цього, розв'язування стохастичної гри можна виконати іншими методами, що ґрунтуються на стохастичній ідентифікації середовища прийняття рішень, наприклад, основаним на законі великих чисел методом оцінювання матриць виграшів (ОМВ) та методом Q-навчання [6–9]. Ці методи вимагають знання структури гри – кількості гравців та кількості їх чистих стратегій. Поточно сформовані матриці виграшів використовуються для побудови векторів змішаних стратегій.

На відміну від ОМВ, метод Q-навчання виконує оцінювання елементів матриць виграшів за ітераційним алгоритмом, причому, у стохастичному формулюванні цей метод забезпечує адаптивне оцінювання – найчастіше будуть враховуватися і обчислюватися елементи матриць, які у середньому забезпечують найбільший виграш.

Об'єктом дослідження цієї роботи є процеси ігрового прийняття рішень в умовах невизначеності.

Предметом дослідження є модель матричної стохастичної гри в умовах невизначеності елементів матриць виграшів.

Метою роботи є побудова мультиагентної моделі стохастичної гри з підкріпленим Q-навчанням [1] для адаптивної ідентифікації матриць виграшів та їх використання для обчислення змішаних стратегій за розподілом Больцмана для підтримки прийняття рішень в умовах невизначеності. Для розв'язування ігрової задачі використано модифікований метод стохастичного Q-навчання.

Модель мультиагентної стохастичної гри

Стохастичну гру агентів у стаціонарному середовищі визначимо кортежем $(I, A^i, V^i \mid \forall i \in I)$, де $I = \{1, 2, ..., L\}$ – множина номерів гравців, $A^i = \{a^i(1), ..., a^i(N_i)\}$ – множина дискретних дій (чисті стратегії) i-го гравця, N_i – кількість стратегій i-го гравця, $V^i : A \to R^i$ – функція винагороди i-го гравця, $A = \underset{i \in I}{\times} A^i$ – множина комбінованих дій гравців, R^i – множина значень виграшів i-го гравця.

Стаціонарне середовище гри задається сукупністю матриць виграшів $\left[v^i(a) \right]_{\forall a \in A}$, $\forall i \in I$, де $v^i(a) = E\{r^i(a)\} = const$ – математичне сподівання випадкових виграшів $\forall a \in A$.

У моменти часу $t = 1, 2, ...$ кожен агент випадково і незалежно від інших вибирає чисті стратегії $a^i \in A^i$. Вибір чистих стратегій здійснюється на основі випадкового розподілу, побудованому на змішаній стратегії $\pi^i = (\pi^i[1], ..., \pi^i[N_i]) \in \Pi^i$, де $\Pi_i = \{\pi_i \mid \sum_{a_i \in A_i} \pi_i(a_i) = 1\}$ – N_i-вимірний одиничний симплекс. Якщо $\pi_i(a_i) \in \{0, 1\}$, то агент здійснює детермінований вибір варіантів рішень.

Після реалізації комбінованого варіанта $a \in A$ гравці отримують випадкові виграші $r^i(a) \in R^i$ з апріорі невідомим розподілом $r^i(a) = Z(v^i(a), d^i(a))$, де $v^i(a)$ – математичне сподівання, $d^i(a)$ – дисперсія. Для моделювання випадкових виграшів приймемо нормальний закон розподілу

$r^i(a) = Normal(v^i(a), d(a))$. Емпірично нормально-розподілені випадкові величини можна отримати за допомогою суми дванадцяти рівномірно-розподілених випадкових чисел $\omega \in [0,1]$:

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

Ігрові ходи кожного агента оцінюються функціями дисконтованих сумарних виграшів:

$$\Upsilon_i = \sum_{t=0}^{\infty}\gamma^t r_t^i \ , \ \forall i \in I \ , \quad (2)$$

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

Дисконтування поточних виграшів здійснюється за законом геометричної прогресії і при $\gamma < 1$ забезпечує швидку стабілізацію функції очікуваного виграшу. Параметр дисконтування γ можна інтерпретувати як ваговий коефіцієнт поточних виграшів, відсоткову ставку, імовірність настання наступного кроку гри тощо. Більший вплив на формування значення функції (2) мають початкові поточні виграші.

Частковим випадком (2) при $\gamma = 1$ є функції середніх виграшів:

$$\Upsilon_n^i = \frac{1}{n}\sum_{t=0}^{n} r_t^i \ , \ n \to \infty \ . \quad (3)$$

На відміну від (2), функція (3) з однаковою вагою враховує усі значення поточних виграшів.

З перерахованих цільових функцій найчастіше у самонавчальних системах прийняття рішень використовується функція (2) з дисконтуванням виграшів. В основному, її використання обумовлено результативністю та легкістю математичних перетворень.

В умовах невизначеності ігрового середовища мета кожного агента полягає у максимізації функції Υ_i за рахунок формування ефективної стратегії π^i:

$$V_\pi^i = E_\pi[\Upsilon_i] \to \max_{\pi_i}, \ \forall i \in I \ , \quad (4)$$

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

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

1) рівноваги за Нешем:

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

2) оптимальності за Парето:

$$V^i(\pi^*) \ge V^i(\pi) \ .$$

У точці рівноваги за Нешем не існує індивідуальної змішаної стратегії i-го гравця ($\forall i \in I$), яка дозволяла б збільшити значення його функції середніх виграшів. У точці оптимальності за Парето не існує колективної змішаної стратегії, яка забезпечувала б збільшення функцій середніх виграшів усіх гравців.

Навчання стохастичної гри

Обчислення функції V_π^i може бути виконано у рекурсивній формі, відомій у літературі як рівняння Беллмана. Враховуючи (1), після нескладних перетворень отримаємо:

$$V_\pi^i(t) = E_\pi(r_t^i) + \gamma\sum_{k=0}^{\infty}\gamma^k E_\pi(r_{t+k+1}^i) = E_\pi(r_t^i) + \gamma V_\pi^i(t+1), \quad (5)$$

де $V_\pi^i(t+1)$ – значення функції дисконтованих виграшів у наступні моменти часу.

Для детермінованої матричної гри поточні значення платіжних функцій можна обчислити так:

$$V_\pi^i(t) = \sum_{a \in A} v^i(a)\prod_{j=1}^{L}\pi_j(a_j) \ .$$

В умовах невизначеності значення елементів $v^i(a)$ матриць виграшів невідомі апріорі. Для їх оцінювання у ході гри використаємо метод ітераційного Q-навчання [6 – 8]:

$$Q_{t+1}^i(a) = (1-\alpha_t)Q_t^i(a) + \alpha_t[r_t^i + \gamma V^i(t+1)], \tag{6}$$

де $a \in A$; α_t – параметр навчання; $V^i(t+1)$ – значення функції виграшів у напрямку оптимального колективного розв'язку.

Вигляд оператора $V^i(t+1)$ у виразі (6) визначається умовою колективної рівноваги, наприклад:

$V^i(t+1) = NE\left(Q_{t+1}^i\right)$ – рівновага за Нешем,

$V^i(t+1) = BR\left(Q_{t+1}^i\right)$ – найкраща відповідь агента,

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

Метод (6) може бути застосований для розв'язування гри одного агента з природою, як часткового випадку N-агентної стохастичної гри при $I = \{i\}, |I| = 1, A = A_i$, коли

$$V^i(t) = \max_{b \in A_i} Q_t^i(b).$$

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 $\pi^{NE} = \left(\pi_1^{NE}, ..., \pi_L^{NE}\right)$ кожному агенту не вигідно відхилитися від власної оптимальної стратегії π_i^{NE}, якщо інші агенти дотримуються точки рівноваги [9]:

$$\sum_{a \in A} Q_t^i(a)\pi_i^{NE}(a_i)\prod_{j \neq i}^{L}\pi_j^{NE}(a_j) \geq \sum_{a \in A} Q_t^i(a)\tilde{\pi}_i(a_i)\prod_{j \neq i}^{L}\pi_j^{NE}(a_j), \tag{7}$$

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

Метод (6) забезпечує виконання умови (7) при

$$NE\left(Q_t^i\right) = \sum_{a \in A} Q_t^i(a)\prod_{j=1}^{L}\pi_j^{NE}(a_j),$$

коли поточне значення оператора вартості станів системи визначається у точці π^{NE} рівноваги за Нешем.

Множина точок NE-рівноваги у змішаних стратегіях є випуклим 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:

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

$$\forall i = 1..N_i,\ \forall a_i \in A_i,\ \pi_i(a_i) > 0,$$

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

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 (6) має вигляд [10]:

$$BR\left(Q_t^i\right) = \max_{\pi_i}\left(\sum_{a \in A} Q_t^i(a)\prod_{j=1}^{L}\pi_j(a_j)\right).$$

Рівновага (оптимальність) за Парето (PE, Pareto Equilibrium) має місце у грі зі спільними інтересами (Common-Interest Markov Game), коли матриці виграшів є однаковими для усіх гравців $Q_t^i(a) = Q_t^j(a) \forall i, j \in I$, $\forall a \in A$ [11].

Гра з різними матрицями виграшів може бути перетворена у гру зі спільними інтересами за допомогою згортки

$$PE\left(Q_t\right)=\sum_{k=1}^{L}\lambda_k\sum_{a\in A}Q_t^k(a)\prod_{j=1}^{L}\pi_j^{PE}(a_j),$$

де $\lambda_j>0$ $(j=1..L)$.

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

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

$$Q_t^i(\pi^{PE})\geq Q_t^i(\pi).$$

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

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

Обчислення значення $V^i(t+1)$ у (6) на основі оптимальних колективних стратегій $\pi^*=\left(\pi_1^*,...,\pi_L^*\right)$ (NE, PE або ін.) вимагає значної обчислювальної роботи. Для практичних застосувань іноді достатньо забезпечити максимізацію платіжних функцій. Тоді, замість (6), можна використати модифікований, подібний до BR, метод оцінювання елементів матриць виграшів за методом стохастичної апроксимації [12]:

$$Q_{t+1}^i(a)=(1-\alpha_t)Q_t^i(a)+\alpha_t(r_t^i+\gamma\max_a Q_t^i(a)). \tag{8}$$

Для забезпечення збіжності методу (8) необхідно накласти обмеження на швидкість зміни його регульованих параметрів. Загальні обмеження є такими:

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

де $\alpha_t>0$ – монотонно спадні додатні послідовності дійсних величин.

На практиці параметр кроку навчання можна обчислити так:

$$\alpha_t=\alpha_0 t^{-\kappa}, \tag{9}$$

де $\alpha_0>0$ – початкове значення параметра α_t; $\kappa\in(0,5;1]$ – порядок швидкості навчання методу (8).

Вибір варіантів рішень в умовах невизначеності здійснюється на основі випадкового розподілу з імовірностями, пропорційними значенням функцій $Q^i(a)$.

Імовірність вибору i-м агентом дії $a_i(k)$ можна визначити на основі розподілу Больцмана:

$$\pi_i(a_i(k))=\frac{e^{Q_i^*(a_i(k))/T}}{\sum_{j=1}^{N_i}e^{Q_i^*(a_i(j))/T}},\ k=1..N_i, \tag{10}$$

де T – температурний параметр системи; $Q_i^*(a_i(k))=\max_{a_{-i}}r^i(a_{-i},a_i(k))$, $a_{-i}\in A_{-i}$, $A_{-i}=\underset{j=1,\ j\neq i}{\overset{L}{\times}}A^j$; $\sum_{k=1}^{N_i}\pi_i(a_i(k))=1$. Для великих значень T реалізується близький до рівномірного випадковий розподіл, а для малих T – розподіл, близький до „жадібного" вибору дій агентів, коли найчастіше вибирається дія з найбільшим Q-значенням.

Вибір чистих стратегій

$$a^i=\left\{A^i(k)\middle|k=\arg\left(\min_k\sum_{j=1}^{k}\pi^i(a^i(j))>\omega\right),\ k=1..N_i\right\}\ \forall i\in I, \tag{11}$$

здійснюється на основі випадкових розподілів, побудованих на змішаних стратегіях $\pi^i=(\pi^i[1],...,\pi^i[N_i])\in\Pi^i\ \ \forall i\in I$, де $\omega\in[0,1]$ – дійсне випадкове число з рівномірним розподілом.

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

$$\Delta = L^{-1}\sum_{i\in I}\left\|\pi_i - \tilde{\pi}_i\right\|^2 , \tag{12}$$

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 π_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 описує розв'язки гри за Нешем у вирівнювальних змішаних стратегіях. Зважувана умова додатково враховує розв'язки гри у чистих стратегіях.

Якісним показником збіжності гри є зростання функцій дисконтованих сумарних виграшів Υ_i (2) або функції виграшів, усередненої за кількістю гравців:

$$\Upsilon = L^{-1}\sum_{i=1}^{L}\Upsilon_i \ . \tag{13}$$

Алгоритм розв'язування стохастичної гри

1. Задати параметри гри: L – кількість гравців; N_i – кількість стратегій кожного гравця $i = 1..L$; $A^i = \{a^i(1),...,a^i(N_i)\}$ – множини дискретних дій (чисті стратегії) гравців; $\left[v^i(a)\right]_{\forall a\in A}$ – матриці виграшів детермінованої гри; початковий момент часу $t = 0$; початкові значення матриць виграшів стохастичної гри: $Q_t^i(a_1,...,a_L) = \varepsilon$, $\forall a_i \in A_i$, $\forall i = 1..L$, де $0 < \varepsilon << 1$ – мале додатне значення та значення параметра дисконтування виграшів $\gamma \in (0,1]$; α_0 – початкове значення параметра α_t; κ – порядок швидкості навчання методу (8); $T > 0$ – температурний параметр системи.

2. Обчислити значення змішаних стратегій $\pi = (\pi_1,...,\pi_L)$ на основі поточних оцінок матриць виграшів $Q_t^i(a_1,...,a_L)$ $\forall i = 1..L$ згідно з (10).

3. Виконати випадковий вибір дій агентів $a = (a_1,...,a_L)$ на основі стратегій $\pi = (\pi_1,...,\pi_L)$ згідно з (11).

4. Отримати поточні виграші агентів $r_t = (r_t^1,...,r_t^L)$ згідно з (1).

5. Обчислити параметр кроку навчання: α_t згідно з (9).

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

7. Обчислити похибку виконання умови доповняльної нежорсткості Δ (12) та значення функції середніх виграшів Υ (13).

8. 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.

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

Результати комп'ютерного моделювання

Виконаємо розв'язування стохастичної гри двох агентів ($L = 2$) з двома чистими стратегіями ($N = 2$) за наведеним вище алгоритмом. Матриці середніх виграшів такої гри подано у табл. 1.

Таблиця 1

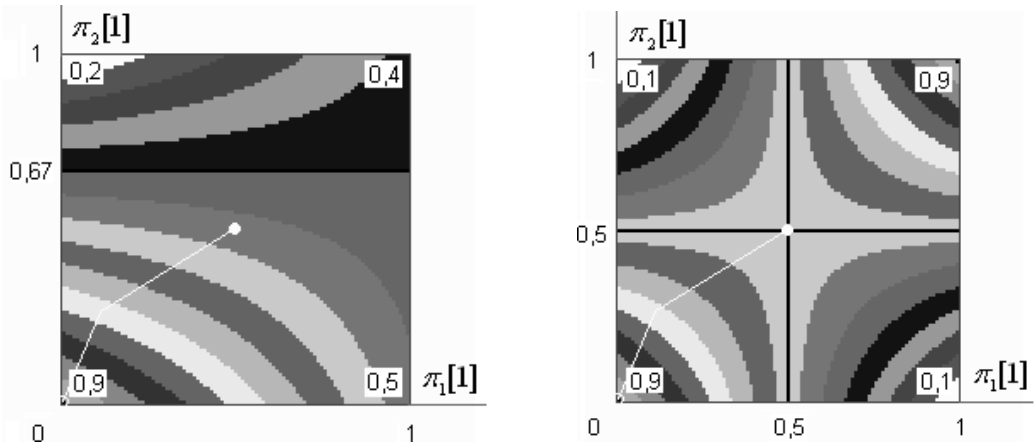
Матриці виграшів гравців

×	Перший гравець		Другий гравець	
Стратегії	$\pi_1(a_1[1])$	$\pi_1(a_1[2])$	$\pi_1(a_1[1])$	$\pi_1(a_1[2])$
$\pi_2(a_2[1])$	0.4	0.2	0.9	0.1
$\pi_2(a_2[2])$	0.5	0.9	0.1	0.9

Дисперсії виграшів приймають однакові значення $d(a) = d = 0,01\ \forall a \in A$ для усіх гравців.

Для розв'язування ігрової задачі виконаємо оцінювання Q-функцій методом (8) з параметрами: $T = 0,1$; $\alpha_0 = 1$; $\kappa = 0,7$; $\gamma = 0,5$. Початкові наближення елементів матриць виграшів приймають однакові значення $Q(a) = 0,01 \quad \forall a \in A$. Змішані стратегії гри визначаються за розподілом Больцмана (10).

Зрізи функцій середніх виграшів V^i, які відповідають даним табл. 1, зображено на рис. 1.



Перший агент *Другий агент*

Рис. 1. Функції середніх виграшів агентів

Як видно на рис. 1, гра має два розв'язки Неша у чистих стратегіях: $(\pi_1[1], \pi_2[1]) = (0;0)$, $(\pi_1[1], \pi_2[1]) = (1;1)$ та один розв'язок у змішаних стратегіях: $(\pi_1[1], \pi_2[1]) = (0,5;0,67)$.

Метод (8) забезпечує розв'язування стохастичної гри у чистих стратегіях – на вершині одиничного симплексу в одній із точок рівноваги за Нешем.

Графіки функцій середніх виграшів Υ та норми відхилення змішаних стратегій від їх цільових значень Δ подано на рис. 2 у логарифмічному масштабі.

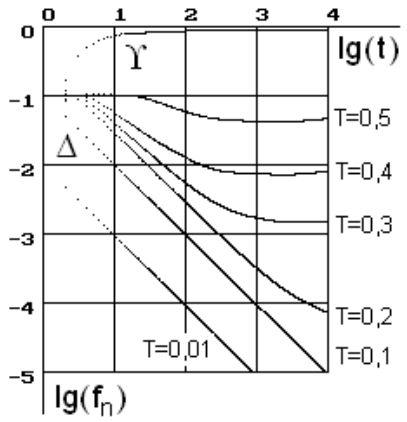


Рис. 2. Характеристики збіжності ігрового Q-методу

Спадання графіка норми похибки виконання умови доповняльної нежорсткості Δ (12) свідчить про збіжність ігрового Q-методу до точки рівноваги за Нешем.

Значення температурного коефіцієнта T справляє значний вплив на збіжність ігрового Q-методу. Швидкість збіжності визначається крутістю спадання графіка функції Δ. Порядок швидкості збіжності можна оцінити тангенсом гострого кута, утвореного між лінійною апроксимацією цього графіка та віссю часу. Зі зростанням значення T порядок швидкості збіжності ігрового Q-методу зменшується.

Як видно на рис. 2, близький до 1 порядок швидкості збіжності методу (8) забезпечується для значень $T \in (0;0,2]$. Розбіжність методу для $T = 0,5$ ілюструється відхиленням траєкторії навчання гри від оптимального значення (0, 0) на рис. 3.

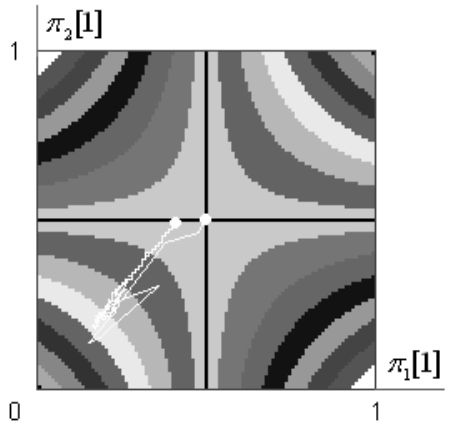


Рис. 3. Відхилення траєкторії гри від оптимального розв'язку для $T = 0,5$

Вивчимо залежність часу збіжності ігрового Q-методу від його параметрів. Необхідний час навчання визначимо як мінімальну кількість кроків, необхідних для досягнення точки рівноваги за Нешем:

$$t_{out} = (t = t_{\min} \mid \Delta_t \le \varepsilon),$$

де ε – задана точність навчання.

Середня кількість кроків навчання агентів визначається у ході проведення ряду експериментів з різними послідовностями випадкових величин:

$$\overline{t} = \frac{1}{k_{\exp}} \sum_{j=1}^{k_{\exp}} t_{out}[j],$$

де $k_{\exp}$ – кількість експериментів.

Результати усереднено за $k_{\exp} = 1000$ експериментами навчання стохастичної гри з точністю $\varepsilon = 10^{-3}$.

На рис. 4 зображено графік залежності середньої кількості кроків навчання стохастичної гри від значення параметра γ дисконтування поточних виграшів (2), отриманий для таких параметрів методу: $\kappa = 0,7$; $T = 0,1$.

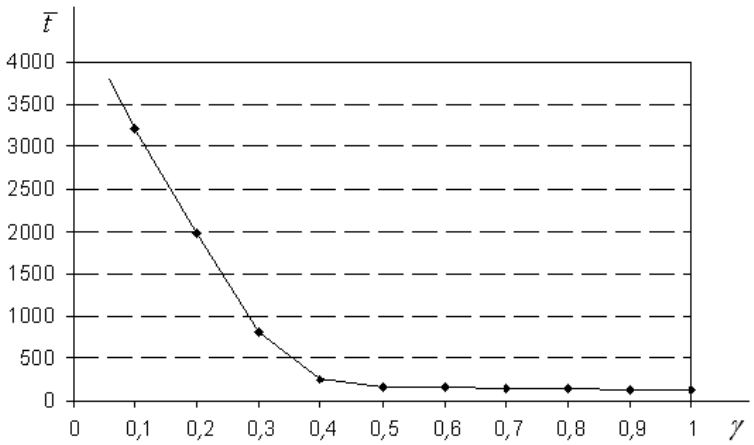


Рис. 4. Залежність середньої кількості кроків навчання стохастичної гри від параметра γ

Зростання γ призводить до сповільнення зменшення дисконтованих поточних виграшів і, відповідно, до зменшення середньої кількості кроків навчання стохастичної гри.

Швидкість збіжності ігрового методу визначається порядком κ швидкості зменшення параметра α_t, який визначає поточну величину кроку навчання методу (8). Залежність середньої кількості кроків навчання стохастичної гри від параметра κ подано на рис. 5. Дані отримано для Q-методу з параметрами: $\gamma = 0,5$; $T = 0,1$.

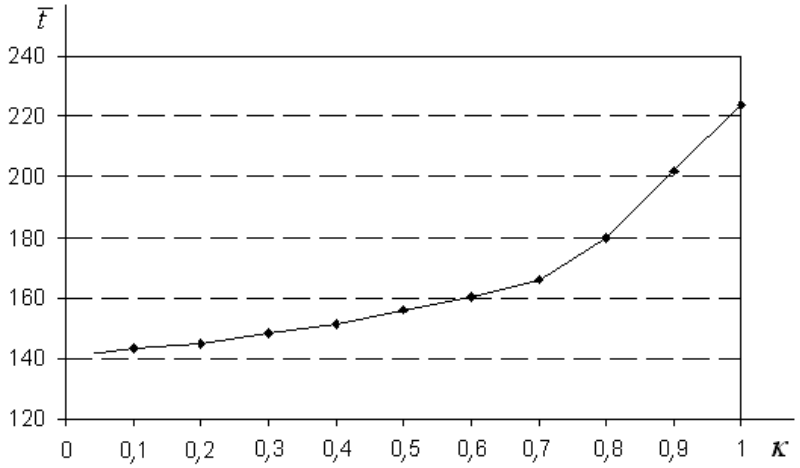


Рис. 5. Залежність середньої кількості кроків навчання стохастичної гри від параметра κ

Із зростанням κ кількість кроків, необхідних для навчання ігрового Q-методу з точністю $\varepsilon = 10^{-3}$, збільшується.

Вплив дисперсії поточних виграшів на збіжність стохастичної гри з Q-навчанням зображено у вигляді графіка на рис. 6.

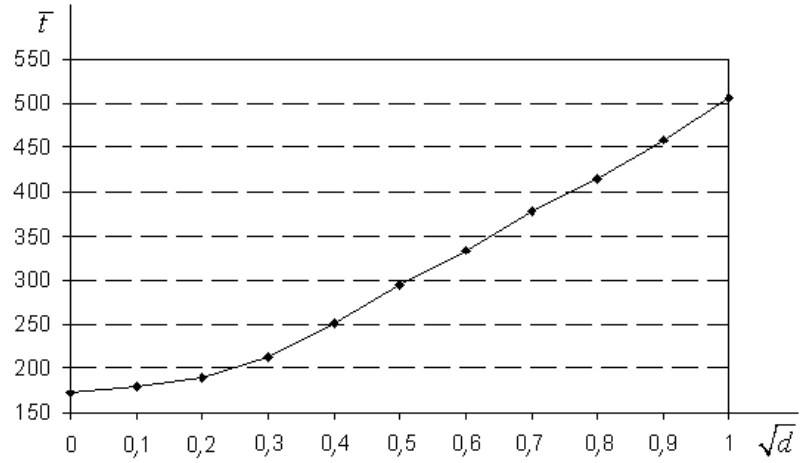


Рис. 6. Залежність середньої кількості кроків навчання стохастичної гри від дисперсії виграшів

Результатом зростання дисперсії d поточних виграшів є збільшення кількості кроків, необхідних для Q-навчання стохастичної гри.

Висновки і перспективи подальших наукових розвідок

Результати проведених досліджень дозволяють стверджувати, що ітераційний метод Q-навчання при дотриманні обмежень на його параметри забезпечує розв'язування стохастичної гри в умовах невизначеності матриць виграшів і може бути використаний для підтримки прийняття колективних рішень.

Практичне використання цього методу обмежується дотриманням умов збіжності до одного із станів колективної рівноваги. В умовах невизначеності гри значення параметрів, які забезпечують виконання умов збіжності, можна встановити теоретично на основі результатів теорії стохастичної апроксимації, або експериментально у ході комп'ютерного моделювання.

У цій роботі експериментально встановлено діапазони зміни параметрів ігрового Q-методу для забезпечення збіжності до однієї із точок рівноваги за Нешем. Із збільшенням значення параметра дисконтування поточних виграшів, зменшенням дисперсії поточних виграшів та зменшенням порядку зміни кроку навчання швидкість збіжності ігрового Q-методу зростає.

1. Доманский В. К. Стохастические игры / В. К. Доманский // Математические вопросы кибернетики. – 1988. – № 1. – С. 26–49. 2. Fudenberg, D. The Theory of Learning in Games / D. Fudenberg, D.K. Levine. – Cambridge, MA: MIT Press, 1998. – 292 p. 3. Weiss G. Multiagent Systems. A Modern Approach to Distributed Artificial Intelligence / G. Weiss, editor. – Springer Verlag, Berlin, 1996. – 643 p. 4. Wooldridge M. An Introduction to Multiagent Systems / M. Wooldridge. – John Wiley & Sons, 2002. – 366 p. 5. Назин, А. В. Адаптивный выбор вариантов: Рекуррентные алгоритмы / А. В. Назин, А. С. Позняк. – М.: Наука, 1986. – 288 с. 6. Watkins, C. J. C. H. Q-Learning / C. J. C. H. Watkins, P. Dayan // Machine Learning. – Kluwer Academic Publishers, Boston. – 1992. – No. 8. – P. 279–292. 7. Kaelbling, Leslie. Reinforcement learning: A survey / Leslie Kaelbling, Michael L. Littman, Andrew W. Moore. Journal of Artificial Intelligence Research. – 1996. – No. 4. – P. 237–285. 8. Sutton, R. S. Reinforcement Learning: An Introduction / Richard S. Sutton, Andrew G. Barto. – MIT Press, 1998. – 322 p. 9. Hu, J. Nash Q-learning for general-sum stochastic games / J. Hu, M. P. Wellman // Machine Learning Research. – 2003. – No. 4. – P. 1039–1069. 10. Weinberg, M. Best-Response Multiagent Learning in Non-Stationary Environments / Michael Weinberg, Jeffrey S. Rosenschein // AAMAS'04. – New York, USA. – July 19–23, 2004. 11. Подиновский В. В. Парето-оптимальные решения многокритериальных задач / В. В. Подиновский, В. Д. Ногин. – М.: Наука, 1982. – 256 с. 12. Граничин О. Н. Введение в методы стохастической аппроксимации и оценивания: учеб. пособие / О. Н. Граничин. – СПб.: Изд-во С.-Петербургского университета, 2003. – 131 с. 13. Мулен Э. Теория игр с примерами из математической экономики / Э. Мулен. – М.: Мир, 1985. – 200 с.