УДК 004.852; 004.94

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

АДАПТИВНІ СТРАТЕГІЇ ПРИЙНЯТТЯ РІШЕНЬ У ГРІ З ПРИРОДОЮ

Досліджується проблема прийняття рішень за допомогою моделі стохастичної гри з природою. Розроблено метод адаптивного вибору варіантів рішень на основі заохочувального навчання та застосування розподілу Больцмана. Розроблено алгоритмічне та програмне забезпечення системи прийняття рішень у грі з природою. Отримано та проаналізовано результати комп’ютерного моделювання стохастичного вибору варіантів рішень.

Ключові слова: прийняття рішень, стохастична гра з природою, адаптивні стратегії.

The problem of decision-making by the instrumentality of model of stochastic game with the nature is investigated. The method of an adaptive choice of decision-making variants on the basis of reinforcement learning and Boltzmann distribution is developed. Algorithmic and software tools of decision-making system in game with the nature are developed. Results of computer modelling of a stochastic choice of decision-making variants are received and analysed.

Keywords: decision-making, stochastic game with the nature, adaptive strategies.

Вступ

Прийняття рішень в умовах ситуативної невизначеності здійснюється за допомогою багатократного повторення дій, які вибираються зі скінченної множини їх можливих варіантів. Необхідність цього зумовлена тим, що в умовах невизначеності пошук цільового рішення не може бути виконаний одноетапним оптимізаційним методом або методом послідовного перебору усіх можливих варіантів рішень та станів системи, оскільки на один і той самий варіант у різні моменти часу система може дати різний сигнальний відгук. Крім цього, повторення дій дає можливість зібрати необхідну інформацію про систему та до деякої міри компенсувати її невизначеність. На практиці важливим є задання способу збереження та опрацювання зібраної інформації.

Для прийняття рішень в умовах невизначеності використовують адаптивні методи [1 – 3], які дають змогу на основі належного опрацювання поточної інформації про систему оптимізувати вибір варіантів дій у наступні моменти часу. Побудова адаптивних методів прийняття рішень може бути виконана на основі математичного апарату випадкових процесів.

Процес одноосібного прийняття рішень можна розглядати у вигляді моделі стохастичної гри з природою. У грі беруть участь інтелектуальний агент та середовище або природа. Агент є моделлю особи, яка приймає рішення. Середовище моделює проблемну галузь прийняття рішень. Гравці включені у контур зворотного зв’язку. У послідовні моменти часу агент вибирає та реалізує один із варіантів дій з відповідним реагуванням на це середовища. Аналізуючи отриманий від середовища сигнал, агент виконує адаптивне корегування стратегії вибору варіантів дій у наступний момент часу. Корегування відбувається так, щоб забезпечити максимальний виграш або мінімальний програш агента у ході всього процесу прийняття рішень.

Перші дослідження стохастичних ігор пов’язані з аналізом ігор автоматів з природою [4–9]. Різні постановки задач залежать від:

1) типу автомата – детермінований або випадковий, з постійною або змінною структурою, з обмеженою або необмеженою кількістю станів, з дискретними або неперервними станами;

2) типу випадкового середовища – стаціонарне, нестаціонарне, з перехідними станами (скінченний або безмежний марківський ланцюг);

3) значення поточних виграшів або програшів – дискретні або неперервні; обмежені або необмежені; скалярні або векторні.

У цих задачах досліджується раціональність поведінки самонавчальних автоматів у випадкових середовищах, яка здебільшого зводиться до асимптотичної мінімізації середнього штрафу за неправильно вибрані дії.

Як математичний апарат дослідження поведінки ігор з природою використовують методи теорії стаціонарних та нестаціонарних ланцюгів Маркова, дифузії, стохастичної апроксимації, мартінгалів, випадкового пошуку, потенціалів, факторизаційних тотожностей, перетворення Лапласа, нелінійного програмування, а також числових методів моделювання на комп'ютері.

Після отримання перших результатів щодо динаміки окремих автоматів розгортаються дослідження ігор групи автоматів у випадкових середовищах як змістовної моделі їх колективної поведінки [5, 9]. Роботи в цьому напрямі підтримував М.Л. Цетлін, який вважав, що складні форми поведінки об'єктів можуть бути реалізовані сукупністю скінченних автоматів, які ведуть себе раціонально у випадкових середовищах.

Ігри декількох скінченних автоматів зі змінною структурою досліджувалися переважно за допомогою моделювання на комп'ютері. Загальний випадок гри N осіб розглянуто у роботі [10], де ігрові методи отримано за допомогою методів стохастичної апроксимації, проекції градієнта та регуляризації.

Стохастичні ігрові моделі здебільшого використовують для розв'язування задач, пов'язаних з необхідністю прийняття рішень в умовах невизначеності – в біології, психології, соціології, політичній науці, військовій справі, економіці, екології, технічних системах тощо.

Незважаючи на значний період розвитку, теорія ігор має нерозв'язані проблеми. Актуальність сучасних досліджень теорії ігор підтверджується, наприклад, великим списком бібліографічних джерел, розміщених на сторінках комп'ютерної мережі Internet. Координує наукові дослідження у галузі теорії ігор Міжнародне товариство динамічних ігор (The International Society of Dynamic Games, адреса в мережі Internet http://www.hut.fi/HUT/Systems.Analysis/isdg/).

Основні проблеми теорії ігор були окреслені на Нобелівському симпозіумі теорії ігор (Nobel Symposium on Game Theory, June 18 – 20, 1993, in Bjorkborn, Sweden) та на низці міжнародних симпозіумів динамічних ігор та їх застосувань (International Symposium on Dynamic Games and Applications).

До визначених перспективних напрямів дослідження ігор належать: повторювані ігри; стохастичні ігри; навчання та адаптація ігор; еволюційні ігри; ігри на переслідування та втікання; мережеві ігри в телекомунікаціях та перевезеннях; динамічні ігри в економіці та менеджменті; ігри у фінансах та маркетингу; керування довкіллям, енергією та ресурсами; торговельні ігри; кооперативні рівноваги ігор; числові методи та комп'ютерні реалізації в ігрових моделях.

Велика увага провідних вчених у теорії ігор спрямована на дослідження стохастичних еволюційних та динамічних повторювальних ігор. На першому за важливістю місці стоять питання побудови та дослідження самонавчальних ігрових методів. Відбувається перехід від пошукових біхевіористичних до когнітивних концепцій побудови самонавчальних ігрових методів. Математичний апарат та засоби моделювання стохастичних ігор використовують як базовий інструмент для дослідження мультиагентних інтелектуальних систем [11]. Продовжується вивчення критеріїв оптимальності колективного співіснування активних елементів (агентів) як в гомогенних, так і в гетерогенних системах розподіленого керування та прийняття рішень.

Стосовно застосувань прикладна теорія ігор перенесла свої початкові пізнавальні інтереси у біології у сферу прагматичніших задач в економіці, менеджменті та маркетингу. Однак потужний математичний апарат теорії ігор поки що не знайшов свого належного місця для вироблення та прийняття рішень під час розв'язування технічних задач.

Відомі методи ігрового вибору варіантів рішень побудовані здебільшого на рекурентних методах зміни змішаних стратегій у межах одиничного симплексу з метою оптимізації функцій середніх виграшів. Інформація про середовище доступна гравцям у вигляді поточних значень виграшів або програшів. У цій роботі пропонується використовувати поточні програші для

адаптивної ідентифікації параметрів середовища прийняття рішень, а змішані стратегії визначати на основі цих параметрів за допомогою розподілу Больцмана.

Метою роботи є побудова ефективного методу прийняття рішень в умовах невизначеності на основі параметричної ідентифікації середовища в адаптивній грі агента з природою.

Формулювання задачі

Розглянемо модель системи прийняття рішень (S, A), яка складається з включених у контур зворотного зв'язку середовища S та агента A [11]. Агент – це активна інтелектуальна система вироблення та реалізації рішень. Взаємодія агента з середовищем описується у термінах гри з природою. Середовище $S = (U, \xi, F)$ задається вектором входів $U = (u(1), u(2), ..., u(N))$, скалярним виходом $\xi \in R^1$ та передатною функцією $F : u \to \xi$. Нехай функція $F = F(v(u), d(u))$ породжується генератором випадкових величин ξ з математичним сподіванням $v(u)$ та дисперсією $d(u)$ $\forall u \in U$. Виходи середовища є оцінками вибраних агентом рішень. Агент $A = (\xi, U, \Pi)$ задається скалярним входом ξ (відповідає виходу середовища), вектором чистих стратегій U (відповідають входам середовища) та правилом $\Pi : \xi \to u$ вибору чистих стратегій. Чисті стратегії визначають дискретні варіанти рішень.

Параметри середовища $v = (v_1, v_2, ..., v_N)$, $d = (d_1, d_2, ..., d_N)$ та вигляд функції розподілу F апріорі не відомі агенту. Вибір варіантів рішень здійснюється у дискретні моменти часу $n = 1, 2, ...$. Після вибору варіанта $u_n = u \in U$ агент спостерігає випадкову реалізацію $\xi_n(u_n) \sim F(v(u_n), d(u_n))$ параметра $v(u_n)$. Значення $\xi_n(u_n)$ інтерпретуватимемо як поточний програш агента за вибір варіанта u_n. Вважається, що випадкові програші $\{\xi_n\}$ є незалежними $\forall u_n \in U$, $n = 1, 2, ...$, мають постійне математичне сподівання $M\{\xi_n(u)\} = v(u) = const$ та обмежений другий момент $\sup_n M\{[\xi_n(u)]^2\} = \sigma^2(u) < \infty$.

Середній програш агента на момент часу *n* визначається так:

$$\Xi_n(\{u_n\}) = \frac{1}{n}\sum_{t=1}^{n} \xi_t \,. \tag{1}$$

Метою агента є мінімізація функції середніх програшів:

$$\overline{\lim_{n\to\infty}} \Xi_n \to \min \,. \tag{2}$$

Для розв'язування задачі (2) необхідно визначити правило Π формування послідовності $\{u_n\}$ варіантів рішень у часі.

Метод розв'язування задачі

Формування послідовності варіантів рішень $\{u_n\}$ з потрібними властивостями виконаємо на основі динамічної змішаної стратегії $p_n = (p_n(1), p_n(2), ..., p_n(N))$. Елементи $p_n(j), j = 1..N$ вектора змішаної стратегії є умовними імовірностями вибору чистих стратегій залежно від поточного варіанта рішення та отриманого програшу. Змішана стратегія набуває значення на одиничному симплексі [10]:

$$S^N = \left\{ p \,|\, \sum_{j=1}^{N} p(j) = 1;\ p(j) \ge 0 \ \ (j = 1..N) \right\}.$$

Правило прийняття рішень

$$\Pi : p_n \xrightarrow{\xi_n} p_{n+1} \xrightarrow{\omega} u_{n+1}$$

повинно забезпечувати переміщення точки змішаної стратегії p_n на одиничному симплексі у середньому напрямку на оптимальний розв'язок під дією поточних програшів ξ_n та вибір варіанта u_{n+1} на основі змішаної стратегії p_{n+1} реалізацією випадкового шансу ω. Ряд рекурентних перетворень вектора змішаної стратегії, які забезпечують розв'язування задачі (2), подано у [10].

У цій роботі пропонується новий рекурентний метод вибору варіантів рішень, що ґрунтується на модифікованому марківському Q-навчанні [12]:

$$Q_{n+1}(u_n) = Q_n(u_n) - \gamma_n \left[\xi_n(u_n) + Q_n(u_n) \right], \tag{3}$$

де $Q_n(u_n) \in R^1$ – функція оцінювання параметра $v(u_n)$ середовища S для вибраного у момент часу n варіанта $u_n = u \in U$; $\gamma_n > 0$ – параметр, що регулює величину кроку методу; $\xi_n(u_n) \in R^1$ – поточний програш агента за вибір варіанта u_n.

Параметр γ_n є монотонно спадною величиною дійсного типу і може бути обчислений так:

$$\gamma_n = \gamma n^{-\alpha}, \tag{4}$$

де $\gamma > 0$; $\alpha > 0$.

Метод (3) здійснює адаптивну параметричну ідентифікацію середовища прийняття рішень залежно від вибраного варіанта u_n та величини поточного програшу ξ_n.

Елементи змішаної стратегії визначаються розподілом Больцмана:

$$p(u) = e^{Q(u)/T} \Big/ \sum_{a \in U} e^{Q(a)/T} \ \forall u \in U, \tag{5}$$

де T – температурний параметр системи.

Перетворення (5) забезпечує належність змішаної стратегії p одиничному симплексу S^N.

Відповідне (5) значення чистої стратегії знаходять з умови:

$$u_n = \left\{ u(k) \,|\, k = \arg\min_k \sum_{i=1}^{k} p_n(i) > \omega \ (k = 1..N) \right\}, \tag{6}$$

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

Збіжність $\lim_{n \to \infty} \left\| p_n - p^* \right\| \to 0$ методу (3) – (6) до оптимальної стратегії p^*, яка мінімізує середні програші (1), забезпечується загальними умовами стохастичної апроксимації [13]:

$$\sum_{n=0}^{\infty} \gamma_n = \infty, \ \sum_{n=0}^{\infty} \gamma_n^2 < \infty.$$

Якість прийняття рішень оцінюється відхиленням $\delta_n = \left\| p_n - p^* \right\|^2$ поточної змішаної стратегії p_n від оптимальної стратегії p^* та усередненим у часі відхиленням:

$$\Delta_n = \frac{1}{n} \sum_{t=1}^{n} \delta_t \ . \tag{7}$$

Алгоритм вибору варіантів рішень

1. Задати початкові значення параметрів:

N – кількість чистих стратегій;

$v = (v_1, ..., v_N)$ – вектор математичних сподівань програшів агента;

$d = (d_1, ..., d_N)$ – вектор дисперсій програшів агента;

$U = (u_1, ..., u_N)$ – вектор значень чистих стратегій;

$p = (1/N, ..., 1/N)$ – змішана стратегія (вектор імовірностей вибору чистих стратегій);

$T > 0$ – температурний параметр розподілу Больцмана;

$Q_0(u) = 0 \forall u \in U$ – ідентифікаційні параметри середовища;

$\gamma > 0$ – параметр кроку навчання;

$\alpha \in (0,1]$ – порядок кроку навчання;

$n_{\max}$ – максимальна кількість кроків методу;

ε – точність обчислення середніх втрат;

$\xi \sim Z(v,d)$ – закон розподілу поточних виграшів;

$n = 0$ – початковий момент часу.

2. Вибрати варіант рішення $u_n \in U$ на основі (6).

3. Отримати значення поточного програшу ξ_n.

4. Знайти значення параметра γ_n (4).

5. Обчислити значення параметра $Q_n(u_n)$ згідно з (3).

6. Визначити елементи вектора змішаної стратегії p_n згідно з (5).

7. Обчислити характеристики якості прийняття рішень δ_n та Δ_n (7).

8. Задати наступний момент часу $n := n+1$.

9. Якщо $n < n_{\max}$ (або $\Delta_n \geq \varepsilon$), то перейти на крок 2, інакше – кінець.

Результати комп'ютерного моделювання

Перевірку працездатності (здатності мінімізувати середні програші) запропонованого методу виконаємо імітаційним програмним моделюванням. Середовище задається $N = 4$ входами $U = (u_1, ..., u_N)$ та одним виходом ξ . Стохастичне перетворення входів на вихідний сигнал здійснюється за допомогою нормального розподілу $\xi \sim Normal(v_i, d_i)$, $i = 1..N$ з векторами апріорі не відомих параметрів $v = (0.5, 0.9, 0.1, 0.7)$ і дисперсій $d = (d_1, ..., d_N)$. Для заданого v оптимальна стратегія $p^* = (0,0,1,0)$ визначає найменший середній програш агента.

Нормально розподілені випадкові величини (за законом Гаусса) обчислюють з використанням суми дванадцяти рівномірно розподілених на відрізку [0,1] величин:

$$\xi_n(u,\omega) = v(u) + \sqrt{d(u)}\left(\sum_{j=1}^{12}\omega_j - 6\right),$$

де $u \in U$; $\omega \in [0,1]$ – дійсне випадкове число з рівномірним законом розподілу.

Бінарні програші $\xi_n(u,\omega) \in \{0,1\}$ знаходять з імовірностями $v(u) \in [0,1]$ $\forall u \in U$:

$$\xi_n(u,\omega) = \begin{cases} 0, & \omega > v(u) \\ 1, & \omega \leq v(u) \end{cases}.$$

У момент часу n агент отримує значення програшу ξ_n та згідно з (3)–(6) реалізує одну із N чистих стратегій. У методі (3) параметр $\gamma_n = \gamma n^{-\alpha}$, де $\gamma = 1$, $\alpha = 0.7$, а початкове значення $Q_0(u) = 0 \forall u \in U$.

Оцінювання асимптотичного порядку швидкості збіжності виконано методом моментів Чжуна [13]:

$$\overline{\lim_{n\to\infty}} n^{\theta} M\{\Delta_n\} \leq \vartheta, \tag{8}$$

де θ – параметр порядку; ϑ – величина швидкості збіжності; Δ_n – усереднена у часі евклідова норма відхилення поточної змішаної стратегії p_n від оптимального значення p^*. Більшому θ та меншому ϑ відповідає більша швидкість збіжності ігрового методу. Довжина досліджуваної вибірки становить 10 тис. кроків.

На підставі оцінки швидкості збіжності (8) поведінка процесу Δ_n у часі апроксимована залежністю $\Delta_n = \vartheta / n^{\theta}$, де $\vartheta > 0$, $\theta \in (0,1]$, $n = 1,2,...$. Після логарифмування отримаємо лінійне співвідношення:

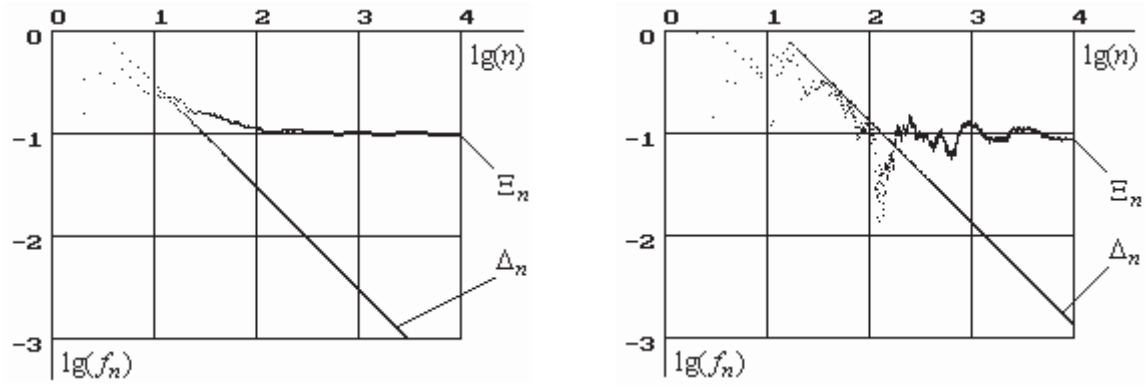
$$\lg \Delta_n = \lg \vartheta - \theta \lg n . \tag{9}$$

З урахуванням цього одержимо залежність $\lg \Delta_n = f(\lg n)$. Тоді параметр $\theta = \lg \Delta_n / \lg n$ вказує на порядок швидкості збіжності досліджуваного методу.

Для визначення порядку швидкості збіжності θ виконана апроксимація випадкового процесу $\lg \Delta_n$ лінійною залежністю (9) на відрізку $\lg n \in [3,4]$ з кроком 0.1 за методом найменших квадратів.

Згладжування випадкової складової швидкості збіжності та виділення порядку цієї швидкості виконується усередненням за реалізаціями випадкового процесу Δ_n .

Результати моделювання подано на рис. 1–4 у логарифмічному масштабі. На рис. 1 зображено графіки функції середніх програшів Ξ_n та функції середньої норми Δ_n відхилення поточної змішаної стратегії від оптимального значення для різних значень дисперсії d нормального розподілу поточних програшів. Дані отримано для значення $T = 0.01$ температурного коефіцієнта розподілу Больцмана.



$d = 0.01$ $d = 1$

Рис. 1. Характеристики збіжності методу

Зменшення у часі функції середніх втрат до мінімального значення $v_3 = 0.1$ та зменшення евклідової норми різниці векторів змішаної та оптимальної стратегій свідчать про працездатність розробленого рекурентного методу. З отриманих результатів видно, що зростання дисперсії призводить до незначного зменшення величини швидкості збіжності ϑ і практично не впливає на порядок θ, який оцінюється тангенсом кута нахилу лінійної апроксимації функції Δ_n з віссю моментів часу. Для заданого T порядок швидкості збіжності θ наближається до 1.

На рис. 2 зображено характеристики збіжності методу (3) – (6) для бінарних програшів при $T = 0.01$. Експериментально встановлено, що зміна закону розподілу поточних програшів (наприклад, на бінарний, рівномірний, експоненційний тощо) значно не впливає на порядок швидкості збіжності розробленого методу.

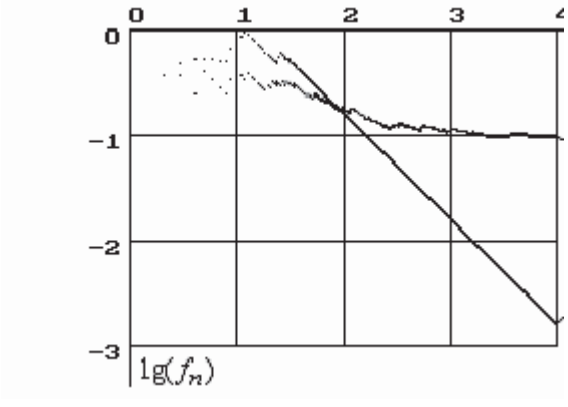


Рис. 2. Характеристики збіжності методу для бінарних програшів

Значний вплив на швидкість збіжності методу (3) – (6) має значення температурного коефіцієнта розподілу Больцмана. На рис. 3 зображено графіки функції похибки Δ_n, отримані для значення дисперсії $d = 0.01$ та різних значень температурного коефіцієнта T .

Зі зменшенням параметра T швидкість збіжності методу (3) – (6) зростає, оскільки за малих T реалізується близький до "жадібного" [14], а за великих T – близький до рівномірного закон вибору варіантів рішень.

Розроблений метод є стійким до стрибкової зміни параметрів середовища, що підтверджується зображеними на рис. 4 результатами. Дані отримано для значень $d = 0.01$, $T = 0.01$. У момент часу $n = 1000$ здійснюється зміна параметра середовища $v_3 = 0.1$ на значення $v_3 = 1$. У момент часу $n = 2000$ відбувається відновлення цього параметра до початкового значення $v_3 = 0.1$.

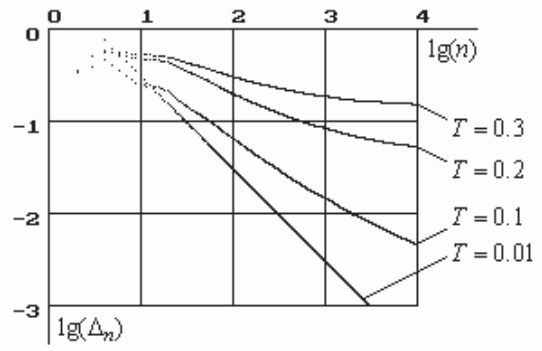


Рис. 3. Вплив температурного коефіцієнта на збіжність методу

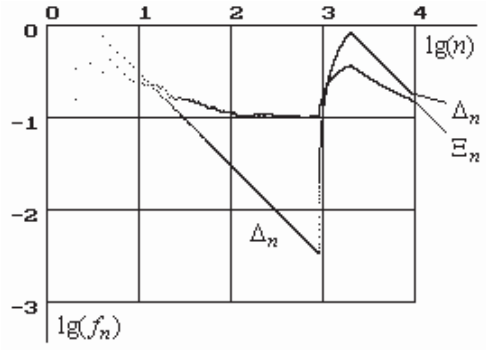


Рис. 4. Ілюстрація стійкості методу

На проміжку $n = 1000...2000$ кроків відбувається перенавчання методу, яке полягає в адаптивному пошуку інших оптимальних стратегій згідно зі зміною параметрів середовища. Після відновлення значення параметра $v_3 = 0.1$ спостерігається збереження початкового порядку швидкості збіжності методу.

Висновки

Розроблений метод прийняття рішень забезпечує стохастичну мінімізацію функції середніх програшів агента і ґрунтується на адаптивній параметричній ідентифікації середовища за допомогою Q -навчання та визначенні змішаної стратегії за розподілом Больцмана. Збіжність методу забезпечується дотриманням базових умов стохастичної апроксимації. Параметри збіжності визначаються теоретично та уточнюються експериментально. За належного підбору параметрів метод забезпечує близький до 1 порядок асимптотичної збіжності, що є граничним значенням для цього класу рекурентних методів.

Розвиток запропонованого методу можливий з урахуванням динаміки станів середовища прийняття рішень та переходу від реактивної до когнітивної моделі агента.

1. Растригин Л.А. Адаптация случайного поиска / Л.А. Растригин, К.К. Рипа, Г.С. Тарасенко. – Рига: Зинатне, 1973. – 242 с. 2. Цыпкин Я.З. Адаптивные методы выбора решений в условиях неопределенности / Я.З. Цыпкин // Автоматика и телемеханика. – 1976. – № 4. – С. 78–91. 3. Срагович В.Г. Теория адаптивных систем / В.Г. Срагович. – М.: Наука, 1976. – 319 с. 4. Robbins H. Some aspects of the sequential design of experiments / H. Robbins // Bulletin of American Mathematical Society. – 1952. – V. 58, No. 5. – P. 527–535. 5. Цетлин, М.Л. Исследования по теории автоматов и моделированию биологических систем / М.Л. Цетлин. – М.: Наука, 1969. – 316 с. 6. Поспелов Д.А. Вероятностные автоматы / Д.А. Поспелов. – М.: Энергия, 1970. – 88 с. 7. Narendra K. Learning Automata: a Survey / K. Narendra., M. Thathachar // IEEE Transactions on Systems, Man and Cybernetics. – 1974. – V. 4. – P. 323–334. 8. Королюк В.С. Автоматы. Блуждания. Игры / В.С. Королюк, А.И. Плетнев, С.Д. Эйдельман // Успехи математических наук. – 1988. – Т. 43, № 1. – С. 87–122. 9. Варшавский В.И. Коллективное поведение автоматов / В.И. Варшавский. – М.: Наука, 1973. – 408 с. 10. Назин А.В. Адаптивный выбор вариантов: Рекуррентные алгоритмы / А.В. Назин, А.С. Позняк. – М.: Наука, 1986. – 288 с. 11. Wooldridge M. An Introduction to Multiagent Systems / M. Wooldridge.– John Wiley & Sons, 2002. – 366 pp. 12. Sutton R. S. Reinforcement Learning: An Introduction / Richard S. Sutton, Andrew G. Barto. – MIT Press, 1998. – 322 pp. 13. Вазан М. Стохастическая аппроксимация / М. Вазан. – М.: Мир, 1972. – 295 с. 14. Кормен Томас Х. Алгоритмы: построение и анализ. – 2-е изд. / Томас Х. Кормен и др. – М.: Вильямс, 2006. – 1296 с.