АКАДЕМИЯ НАУК СССР

ИЗВЕСТИЯ АКАДЕМИИ НАУК СССР

ТЕХНИЧЕСКАЯ КИБЕРНЕТИКА

(ОТДЕЛЬНЫЙ ОТТИСК)

2

МОСКВА 1987

ТЕХНИЧЕСКАЯ КИБЕРНЕТИКА
№ 2 · 1987

УДК 62–50:519.217.2

АДАПТИВНЫЙ АЛГОРИТМ НАХОЖДЕНИЯ СЛАБО ЭФФЕКТИВНОГО ВАРИАНТА В УСЛОВИЯХ СЛУЧАЙНОСТИ

СУЧАНСКИЙ М. Е.

Введение. Основополагающая работа М. Л. Цетлина о выборе автоматом с постоянной структурой оптимального действия в стационарной случайной среде привела к появлению ряда исследований коллективного поведения автоматов [1, 2]. Коллективное поведение автоматов в них рассматривалось как игра нескольких автоматов, каждый из которых действует на основе единственного критерия оптимальности, причем платежные функции им неизвестны. В [3–5] был предложен метод анализа асимптотического поведения ε-оптимальных автоматов в играх и было показано, что для всех построенных конструкций автоматов характерно разыгрывание максиминных, а не равновесных по Нэшу партий. С точки зрения теории адаптивных систем [6], игры автоматов описываются как управление векторным случайным процессом с независимыми компонентами с помощью прямого произведения ε-оптимальных автоматов. В последнее десятилетие активизировались исследования обучающихся автоматов с переменной структурой [7–9]. Игра таких автоматов обычно интерпретируется как процесс децентрализованного принятия решения в условиях априорной неопределенности. В ходе игры сконструированных на основе метода стохастической аппроксимации обучающихся автоматов удается находить равновесное по Нэшу решение [7, 10]. Однако довольно часто встречаются задачи, в которых требуется найти Парето-оптимальный вариант (управление) в условиях случайности вектора величин, характеризующих выбираемый вариант, и отсутствия информации о соответствующих функциях распределения. Существующие методы [11, 12] не позволяют подойти к решению этой проблемы. В данной работе строится ε-оптимальная последовательность многокритериальных автоматов $A^{(n)}$. Доказывается, что с ростом числа состояний памяти в этом автомате стремится к единице финальная вероятность выбора оптимального по Слейтеру управления (варианта) в соответствующей марковской цепи. Помимо технических приложений данный подход может быть полезен для построения моделей коллективного взаимодействия автоматов, действующих в условиях многих взаимосвязанных критериев [13–15].

1. Постановка задачи. Сформулируем задачу, придерживаясь терминологии и обозначений из [6]. Пусть имеется векторный управляемый случайный процесс $\xi_t=(\xi_t^{(1)}, \ldots, \xi_t^{(k)})$ с дискретным временем $t=1, 2, \ldots$ и векторное пространство управлений $Y=Y^1\times\ldots\times Y^m$, т. е. управлением в момент t служит вектор $\mathbf{y}_t=(y_t^1, \ldots, y_t^m)$. Будем считать, что размерность вектора ξ_t — целое число $k\geqslant 2$, а размерность вектора $\mathbf{y}_t$ — целое число $m\geqslant 1$. Относительно управляемого процесса с измеримым фазовым пространством $(X, \mathscr{X})$ будем предполагать, что: а) каждая компонента процесса $\xi_t^{(i)}$ $(i=\overline{1, k})$ принимает значения из конечного отрезка $[a, b]$ (пусть для определенности $a=0$, $b=1$); б) все множества Y^j $(j=\overline{1, m})$ конечны и в множестве Y^j содержится $\varkappa_j$ элементов, т. е. всего имеется $\varkappa=\varkappa_1\cdot\ldots\cdot\varkappa_m$ различных управлений (вариантов);

в) управляемый процесс принадлежит классу однородных процессов с независимыми значениями (ОПНЗ). Это значит, что при каждом $M \in \mathscr{X}$ его распределение вероятностей ($M = M_1 \times \ldots \times M_k$)

$$\mu(M \mid \mathbf{y}_{t-1}) = P(\xi_t^{(1)} \in M_1, \ldots, \xi_t^{(k)} \in M_k \mid \mathbf{y}_{t-1}) \tag{1.1}$$

не меняется во времени (однородность), зависит только от последнего управления $\mathbf{y}_{t-1}$ и не зависит от предыдущих значений процесса и управлений; г) математические ожидания компонент вектора ξ_t:

$$W_i(\mathbf{y}) = \int_X x_i \mu(dx_1, \ldots, dx_i, \ldots, dx_k \mid \mathbf{y})$$

таковы, что

$$0 < W_i(\mathbf{y}) < 1 \quad \text{при всех} \quad i = \overline{1, k} \quad \text{и} \quad \mathbf{y} \in Y. \tag{1.2}$$

Будем их интерпретировать как средние выигрыши при выборе управления (варианта) $\mathbf{y}$. Далее будем рассматривать только ОПНЗ ξ_t с независимыми компонентами, т. е.

$$\mu(M \mid \mathbf{y}) = \prod_{i=1}^{k} \mu_i(M_i \mid \mathbf{y}).$$

Для таких процессов математические ожидания компонент равны

$$W_i(\mathbf{y}) = \int_0^1 x \mu_i(dx \mid \mathbf{y}). \tag{1.3}$$

Рассмотрим условное распределение вероятностей на Y $H_t = H_t(N \mid \xi_t, \mathbf{y}_{t-1})$, $N \subset Y$, $t \geqslant 1$, которое называется правилом выбора управления в момент t. Это правило использует значение управляемого процесса ξ_t в данный момент времени и управление $\mathbf{y}_{t-1}$ в предшествующий момент времени и, возможно, является вероятностным, т. е. имеет невырожденное распределение. Алгоритмом (стратегией) управления A называется совокупность правил выбора управления в каждый момент времени t, т. е. $A = \{H_t, t = 1, 2, \ldots\}$.

Сформулируем цель управления в терминах функций $W_i(\mathbf{y})$ из формулы (1.3), $i = \overline{1, k}$. Поскольку $W_i(\mathbf{y})$ содержательно означает i-ю компоненту вектора выигрышей, то естественно стремиться найти такое $\mathbf{y}_0$, которому отвечает неулучшаемый по всем компонентам одновременно вектор $\mathbf{W}^0 = \mathbf{W}(\mathbf{y}_0)$. Напомним, что вектор $\mathbf{W}^0$ называется оптимальным по Слейтеру (слабо эффективным), если не существует среди всех возможных векторов $\mathbf{W}$ такого, что $W_i > W_i^0$ для всех $i = \overline{1, k}$ [12]. Согласно теореме Гермейера [11, 12], при выполнении условия (1.2) вектор $\mathbf{W}^0$ слабо эффективен тогда и только тогда, когда существует вектор ρ, удовлетворяющий

$$\rho_i > 0, \quad \sum_{i=1}^{k} \rho_i = 1$$

такой, что

$$\min_{1 \leqslant i \leqslant k} \rho_i W_i^0 = \max_{\mathbf{y} \in Y} \min_{1 \leqslant i \leqslant k} \rho_i W_i(\mathbf{y})$$

В данной работе предполагается, что априорно задан вектор ρ, удовлетворяющий сформулированным ограничениям, компоненты которого называются весами или коэффициентами важности. Пусть $\mathbf{W}^0$ — это такой оптимальный по Слейтеру вектор, который соответствует заданному вектору весов ρ. Управление $\mathbf{y}_0$, отвечающее $\mathbf{W}^0 = \mathbf{W}(\mathbf{y}_0)$, будем называть слейтеровским, а множество всех таких управлений обозначим Y_c. Множество Y_c непусто в силу конечности Y. Подчеркнем, что если бы имелась

априорная информация о функциях $W_i(\mathbf{y})$, то задача нахождения множества оптимальных по Слейтеру вариантов решалась бы традиционными неадаптивными методами. В условиях отсутствия такой информации задача состоит в синтезе алгоритма управления A_ε такого, чтобы для всех управляемых случайных процессов ξ_t описанного выше класса при заданных весах ρ и $\varepsilon>0$

$$\lim_{t\to\infty} \min_{1\leq i\leq k} W_i(t, A_\varepsilon) > \min_{1\leq i\leq k} \rho_i W_i^0 - \varepsilon. \tag{1.4}$$

Здесь $W_i(t, A_\varepsilon)$ — i-я компонента среднего выигрыша в момент t по мере, порождаемой алгоритмом управления A_ε. Заметим, что коэффициент ρ_i учитывается в самой величине $W_i(t, A_\varepsilon)$ при переходе к бинарной версии ОПНЗ. Это отражено в формуле (4.1). Заметим, что неопределенность, возникающую из-за наличия нескольких показателей $W_i(\mathbf{y})$, по которым оценивается выбираемое управление $\mathbf{y}$, можно преодолевать не только с помощью свертки $\min \rho_i W_i(\mathbf{y})$, но и другими способами. Выбор той или иной свертки показателей определяется априорной информацией о практически решаемой многокритериальной задаче.

2. Прикладное значение задачи. Сформулированная выше задача встречается на практике. Пусть имеется вычислительная система (ВС) определенной структуры и программа, функциональные операторы которой представлены в виде ярусно-параллельного графа [16]. Программу необходимо реализовать на ВС, распределив определенным образом ее операторы по машинам ВС. Если данная программа используется неоднократно, то возникает проблема найти такое распределение ее операторов, чтобы минимизировать общее время выполнения программы. Для того, чтобы точнее сформулировать задачу, введем ряд допущений. Будем рассматривать синхронный режим или случай полной функциональной связности ярусов. В этом случае выполнение операторов следующего яруса начинается только после реализации всех операторов данного яруса. Будем предполагать, что ВС неоднородна, т. е. входящие в нее машины имеют разную производительность. Из сказанного ясно, что достаточно ограничиться задачей распределения операторов одного яруса на систему параллельно работающих машин. Тогда в нашем распоряжении имеется d операторов, которые можно параллельно выполнять на d машинах ВС: по одному оператору на одной машине. Всякое распределение операторов по машинам будем называть вариантом. Очевидно, что всего имеется $\varkappa_1=d!$ вариантов, а размерность Y равна $m=1$. Пусть $\tau_i(y)$, $i=\overline{1, d}$,— случайная величина, равная времени выполнения оператора на i-й машине при выборе варианта распределения y. Случайность времени появляется из-за того, что длительность выполнения оператора может зависеть от конкретных значений данных, от реального количества осуществляемых циклов и т. д. У величины $\tau_i(y)$ имеется неизвестное нам конечное математическое ожидание $t_i(y)=E\tau_i(y)$. В данном случае средние времена $t_i(y)$ имеют смысл проигрышей, но с учетом этого различия аналогичны средним выигрышам $W_i(y)$. Поскольку продолжительность выполнения яруса определяется в синхронном режиме наибольшим среди всех машин временем $\tau_i(y)$ $(i=\overline{1, d})$, то задача состоит в построении адаптивного алгоритма нахождения такого варианта распределения y_l $(l=\overline{1, d!})$, для которого достигается

$$\min_{y_l} \max_{1\leq i\leq d} t_i(y_l).$$

Данная задача есть частный случай сформулированной ранее при $\rho_i=1/d$, $i=\overline{1, d}$.

Известно, что задача о целесообразном поведении автомата в стационарной случайной среде была сформулирована М. Л. Цетлиным как модель поведения некоего животного в Т-образном лабиринте [14]. Эксперименты, послужившие основой для этой модели, состояли в том, что живот-

ное, привлекаемое запахом пищи, поворачивало либо в левый, либо в правый коридор и получало там удары током с вероятностями p_l и p_r. Целесообразность поведения животного состояла в том, что после нескольких попыток, оно использовало лишь тот коридор, в котором удары током бывали реже. Вообразим теперь более сложную ситуацию. Пусть в лабиринте имеется $\varkappa_1$ коридоров, в каждом из которых с вероятностью p_i' бьют током, а пища достается животному теперь не всегда, а с вероятностью p_i'' $(i=\overline{1, \varkappa_1})$. Можно предположить, что животное после нескольких попыток сумеет выбрать коридор, в котором реже бьют током и чаще дают пищу. Алгоритм выбора ε-оптимального по Слейтеру варианта должен имитировать это довольно сложное поведение.

3. Алгоритм управления. Поскольку компоненты ОПНЗ ξ_t независимы, то естественно стремиться синтезировать управляющий алгоритм A_ε настолько децентрализованным, насколько это возможно.

Как показывает анализ некоторых классов игр автоматов (при этом $k=m$), прямое произведение ε-оптимальных автоматов при глубине памяти $n\to\infty$ с финальной вероятностью 1 выбирает в них слейтеровские управления [4, 6], в случае $\rho_i=1/k$ для всех $i=\overline{1, k}$. Принцип «субъективной удовлетворенности», сформулированный в [5], показывает, что поведение такого типа присуще автоматам из ε-оптимальных последовательностей. Однако построен пример игры двух автоматов, в которой с ростом глубины памяти n растет финальная вероятность неслейтеровского управления [3]. Поэтому приходится отказаться от децентрализации управления векторным ОПНЗ и вводить некоторую координацию при смене старого управления на новое. Предлагаемый ниже алгоритм навеян соображениями И. М. Гельфанда и М. Л. Цетлина о простейшем методе нелокального поиска: «сочетание принципа гомеостата с каким-либо локальным методом» [1]. Говоря нестрого, стратегия A_ε занимает промежуточное положение между централизованными и децентрализованными системами управления [14]. Оказывается, что анализ работы алгоритма проводится методом, предложенным в [4].

Стратегию управления A_ε образуют преобразователь входа и основная часть, за которой сохраним наименование «алгоритм управления A_ε». Преобразователь входа осуществляет квантование значений процесса ξ_t и переход к его бинарной версии [2, 6]. Если i-я компонента процесса в момент t принимает значение $x_i\in[0, 1]$, ставим ей в соответствие число $z=1$ с вероятностью $\rho_i x_i$ и число $z=0$ с вероятностью $1-\rho_i x_i$. Тогда полученная бинарная версия ОПНЗ ξ_t описывается условными вероятностями

$$q_i(\mathrm{y})=P\{z=1|\mathrm{y}\}=\rho_i W_i(\mathrm{y}),$$

$$p_i(\mathrm{y})=P\{z=0|\mathrm{y}\}=1-\rho_i W_i(\mathrm{y}).$$

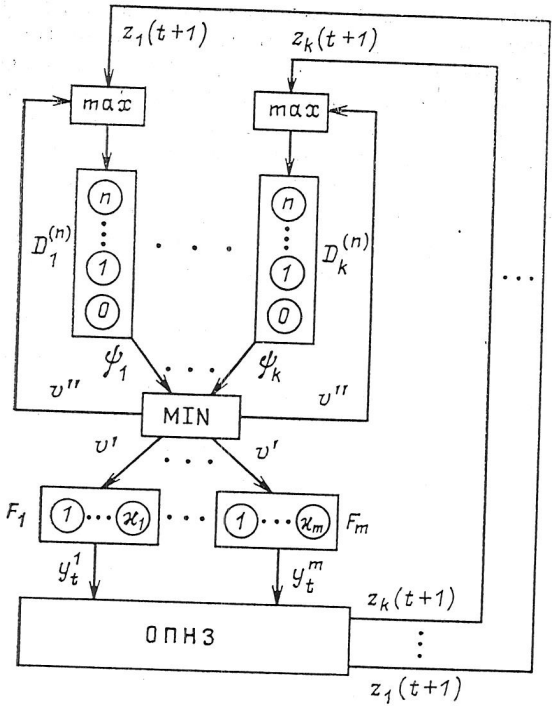
Значение $z=1$ часто называют поощрением, а значение $z=0$ — штрафом. Поскольку все $\rho_i>0$ $(i=\overline{1, k})$ и в силу условия (1.2) имеем:

$$0<q_i(\mathrm{y})<1 \quad \text{для всех} \quad i=\overline{1, k} \quad \text{и} \quad \mathrm{y}\in Y. \tag{3.1}$$

Система управления A_ε состоит из почти автономных подсистем, автономия которых не сохраняется лишь в некоторые моменты времени. Эти моменты нарушения децентрализованного характера управления случаются всякий раз, когда хотя бы одна подсистема готовится сменить свое действие. Такая подсистема информирует о своих намерениях все остальные, что приводит к выбору нового действия каждой из подсистем, после чего возобновляется автономный способ управления. Алгоритм управления A_ε опишем как автомат, полученный в результате композиции нескольких подавтоматов (см. рисунок).

Рассмотрим сначала подавтоматы памяти

$$D_l^{(n)}=\{Z, S_l^{(n)}, \Psi_l, B^{(n)}\}, \quad l=\overline{1, k},$$



где $Z=\{0;\ 1\}$ — множество входных воздействий; $S_l^{(n)}=\{s_{\lambda_l}\}$ — множество состояний из $(n+1)$ элемента, нижний индекс $\lambda_l\in\overline{0,\ n}$; $\Psi_l=\{-1;\ +1\}$ — двухэлементное множество выходных сигналов; $B^{(n)}=\{B_0^{(n)},\ B_1^{(n)}\}$ — две стохастические матрицы переходов между состояниями: $B_0^{(n)}=\|b_{ij}(0)\|$ — матрица переходов при получении подавтоматом $D_l^{(n)}$ штрафа, а $B_1^{(n)}=\|b_{ij}(1)\|$ — при получении поощрения. Матрицы $B^{(n)}$ одинаковы у всех подавтоматов $D_l^{(n)}$. В каждой строке этих матриц имеется единственный элемент, отличный от 0 и равный 1. Переходы между состояниями у подавтомата $D_l^{(n)}$ происходят так же, как и в лепестке «доверчивого» автомата Роббинса — Кринского [1, 6, 17]. А именно, получив поощрение, автомат из любого состояния переходит в состояние с индексом $\lambda_l=n$. Получив штраф, он переходит в состояние с индексом на единицу меньше, за исключением состояния с индексом $\lambda_l=0$. Как будет ясно из дальнейшего, автомат A_ε устроен так, что в состоянии с индексом $\lambda_l=0$ подавтомат $D_l^{(n)}$ может получить только поощрение. Функция выходов: в состоянии с индексом $\lambda_l=0$ выход $D_l^{(n)}$ равен $\psi_l=-1$, а во всех остальных состояниях выход равен $\psi_l=+1$. Индекс λ_l состояния s_{λ_l} будем также называть глубиной этого состояния, а глубиной памяти подавтомата $D_l^{(n)}$ (и автомата A_ε) — максимальную глубину состояния. Автомат A_ε с глубиной памяти n будем обозначать $A_\varepsilon^{(n)}$.

Подавтомат MIN имеет единственное состояние. Множеством его входных воздействий служит прямое произведение множеств Ψ_l, $l=\overline{1,\ k}$, т. е. на каждом такте работы он получает на входе вектор $(\psi_1,\ldots,\ \psi_k)$, где $\psi_l\in\{-1;\ +1\}$. У подавтомата MIN имеется $(k+m)$ выходных каналов. Выходной сигнал в каждом из m каналов, который является входным для одного из подавтоматов F_r $(r=\overline{1,\ m})$, равен $v'=\min\limits_{1\leqslant l\leqslant k}\psi_l$. Выходные сигналы в остальных k каналах равны $v''=-\min\limits_{1\leqslant l\leqslant k}\psi_l$. Максимум из сигнала

в l-ом из этих каналов и величины z_l поступит на вход подавтомата $D_l^{(n)}$ $(l=\overline{1, k})$ в следующий момент времени.

Подавтоматы F_r $(r=\overline{1, m})$ функционируют одинаково, но отличаются по числу внутренних состояний: $F_r=\{V, S_r^{(\varkappa_r)}, Y_r, U_r\}$. Множество входных воздействий $V=\{-1; +1\}$. Множество состояний $S_r^{(\varkappa_r)}=\{s^{\gamma_r}\}$ содержит $\varkappa_r$ элементов, где верхний индекс $\gamma_r\in\overline{1, \varkappa_r}$; $U_r=\{U_r(-), U_r(+)\}$ — две стохастические матрицы переходов между состояниями. Матрица $U_r(-)=\|u_{ij}^{(-)}\|$ соответствует переходам при входном сигнале $v'=-1$, причем все элементы каждой строки матрицы $U_r(-)$ одинаковы и равны $1/\varkappa_r$. Матрица $U_r(+)=\|u_{ij}^{(+)}\|$ соответствует переходам при входном сигнале $v'=+1$, причем $u_{ii}^{(+)}=1$, $u_{ij}^{(+)}=0$ при $i\neq j$. Множество выходных сигналов $Y^r=\{y^{\gamma_r}\}$, где $\gamma_r\in\overline{1, \varkappa_r}$, является множеством значений r-й компоненты вектора управлений. Множество Y^r будем называть множеством действий r-го подавтомата F_r, а элементы этого множества — действиями. Взаимно-однозначная функция выходов подавтомата F_r сопоставляет каждому состоянию s^{γ_r} из $S_r^{(\varkappa_r)}$ элемент y^{γ_r} из множества Y^r. Из сказанного ясно, что подавтомат F_r функционирует следующим образом. При получении $v'=+1$ он сохраняет состояние, а следовательно, и действие, а при получении $v'=-1$ он равновероятно выбирает новое состояние (действие) среди всех возможных (разрешается возврат в исходное состояние).

Функционирование автомата $A_\varepsilon^{(n)}$ ясно из рисунка. У него имеется k внешних входных узлов и m внешних выходных узлов. Выходом автомата $A_\varepsilon^{(n)}$ на такте t будет вектор управления $\mathbf{y}_\gamma=\{y_t^{\gamma_1}, \ldots, y_t^{\gamma_m}\}$, компоненты которого независимо вырабатываются подавтоматами F_r, $r=\overline{1, m}$. ОПНЗ ξ_t некоторым образом отреагирует на управление $\mathbf{y}_\gamma$. После приведения полученной реакции к бинарному виду (наказание или поощрение), на вход автомата $A_\varepsilon^{(n)}$ поступит на такте $t+1$ вектор $\mathbf{z}(t+1)=(z_1(t+1), \ldots, z_k(t+1))$, где $z_l\in\{0; 1\}$, $l=\overline{1, k}$. Взаимодействие автомата $A_\varepsilon^{(n)}$ и векторного ОПНЗ описывается марковской цепью $\mathrm{M}_n=(S^{(n)}, P^{(n)})$. Ее множеством состояний служит $S^{(n)}=S_1^{(n)}\times\ldots\times S_k^{(n)}\times S_1^{(\varkappa_1)}\times\ldots\times S_m^{(\varkappa_m)}$ прямое произведение множеств состояний подавтоматов $D_1^n, \ldots, D_k^{(n)}$, $F_1, \ldots, F_m$. Состоянию $\mathbf{s}=(\mathbf{s}_\lambda, \mathbf{s}^\gamma)=(s_{\lambda_1}, \ldots, s_{\lambda_k}, s^{\gamma_1}, \ldots, s^{\gamma_m})$ из $S^{(n)}$ отвечает ситуация, когда автоматы памяти $D_l^{(n)}$ $(l\in\overline{1, k})$ находятся в состояниях глубины λ_l $(\lambda_l=\overline{0, n})$, а автоматы действия F_r находятся в состояниях s^{γ_r} и совершают действие y^{γ_r} $(\gamma_r=\overline{1, \varkappa_r})$. Множество $S^{(n)}$ состоит из $(n+1)^k\varkappa$ элементов. Пусть $(\mathbf{s}_\lambda, \mathbf{s}^\gamma)\in S^{(n)}$ и $(\mathbf{s}_\mu, \mathbf{s}^\delta)\in S^{(n)}$. В дальнейшем будем обозначать состояния цепи их индексами, т. е. $\mathbf{s}=(\mathbf{s}_\lambda, \mathbf{s}^\gamma)$ как (λ, γ), а вероятность перехода между состояниями (λ, γ) и (μ, δ) как $p\langle\lambda, \gamma|\mu, \delta\rangle$. Из сказанного выше следует, что элементы матрицы переходных вероятностей $P^{(n)}=\|p\langle\lambda, \gamma|\mu, \delta\rangle\|$ вычисляются следующим образом [18]. Пусть для всех $l=\overline{1, k}$ глубина $\lambda_l>1$. Тогда: $p\langle\lambda, \gamma|\mu, \delta\rangle=0$ при $\gamma\neq\delta$,

$$p\langle\lambda, \gamma|\mu, \gamma\rangle=\prod_{i=1}^{k}[q_i(\mathbf{y}_\gamma)b_{\lambda_i\mu_i}(1)+p_i(\mathbf{y}_\gamma)b_{\lambda_i\mu_i}(0)], \tag{3.2}$$

где $\mathbf{y}_\gamma$ — выходной вектор автомата $A_\varepsilon^{(n)}$ в состоянии (λ, γ). Пусть имеется $G=\{i(1), \ldots, i(g)\}$ — непустое подмножество $\{1, \ldots, k\}$ такое, что $\lambda_{i(1)}=1, \ldots, \lambda_{i(g)}=1$, а для всех остальных $l\in\{1, \ldots, k\}$ глубина состояния автомата памяти $D_l^{(n)}$ $\lambda_l>1$. Тогда

$$p\langle\lambda, \gamma|\mu, \delta\rangle=\prod_{i\notin G}[q_i(\mathbf{y}_\gamma)b_{\lambda_i\mu_i}(1)+p_i(\mathbf{y}_\gamma)b_{\lambda_i\mu_i}(0)].$$

$$\cdot\Big[\prod_{i\in G} q_i(\mathbf{y}_\gamma)\, b_{\lambda_i\mu_i}(1)\cdot \prod_{r=1}^{m} u^{(+)}_{\gamma_r\delta_r} + \Big(\prod_{i\in G}[p_i(\mathbf{y}_\gamma)\, b_{\lambda_i\mu_i}(0) + q_i(\mathbf{y}_\gamma)\, b_{\lambda_i\mu_i}(1)] -$$

$$- \prod_{i\in G} q_i(\mathbf{y}_\gamma)\, b_{\lambda_i\mu_i}(1)\Big)\cdot \prod_{r=1}^{m} u^{(-)}_{\gamma_r\delta_r}\Big]. \tag{3.3}$$

Пусть имеется $l\in\{1,\ldots,k\}$ такое, что $\lambda_l=0$. Пусть также $(n,\gamma)=(s_n,\ldots,s_n,s^{\gamma_1},\ldots,s^{\gamma_m})$ — состояние цепи, соответствующее случаю, когда все автоматы памяти находятся в самом глубоком состоянии. Тогда

$$p\langle\lambda,\gamma|n,\gamma\rangle=1. \tag{3.4}$$

Матрица $P^{(n)}=\|p\langle\lambda,\gamma|\mu,\delta\rangle\|$ — стохастическая, т. е. вероятности перехода (3.2)—(3.4) действительно определяют марковскую цепь. Соотношения (3.2)—(3.4) имеют простой смысл. Если все автоматы памяти $D_1^{(n)},\ldots,D_k^{(n)}$ находятся в состояниях глубины больше 1, то ни один из них в данный момент времени не может сменить выходной сигнал ψ_l $(l=\overline{1,k})$ и, следовательно, автоматы действия сохраняют вектор управления y_γ. В этом случае переходы в цепи M_n происходят так же, как при управлении процессом ξ_t с помощью прямого произведения автоматов Роббинса — Кринского и описываются вероятностями (3.2). Если же хотя бы один автомат памяти $D_l^{(n)}$ находится в состоянии глубины 1, т. е. $G\neq\varnothing$, и получает штраф, то он переходит в состояние глубины 0, и все автоматы действия F_r $(r=\overline{1,m})$ независимо выбирают новое действие равновероятным образом. Описанный выше автомат $A_\varepsilon^{(n)}$ устроен так, что при попадании марковской цепи в состояние (λ,γ), соответствующее состоянию глубины 0 хотя бы у одного автомата памяти, согласно (3.4), в следующий момент времени с вероятностью 1 происходит переход в состояние (n,γ). В дальнейшем такие состояния цепи, которым отвечают состояния глубины n всех автоматов памяти, будем называть глубокими. Итак, выбранный алгоритм управления $A_\varepsilon^{(n)}$ существенно отличается от стратегии, реализуемой прямым произведением автоматов Роббинса — Кринского, наличием координации в моменты смены управления.

4. Сходимость алгоритма. Пусть $\mathbf{y}=(y^{\gamma_1},\ldots,y^{\gamma_m})$ — некоторое управление, $\gamma_r\in\overline{1,\varkappa_r}$. Обозначим $S^{(n)}(\mathbf{y})$ — множество состояний цепи, каждому из которых отвечает управление $\mathbf{y}$, т. е. $S^{(n)}(\mathbf{y})=\{\mathbf{s}|\mathbf{s}=(\mathbf{s}_\lambda,\mathbf{s}^\gamma)$ для всех $\lambda=(\lambda_1,\ldots,\lambda_k)$, где $\lambda_l=\overline{0,n}$ $(l=\overline{1,k})\}$. Подмножества $S^{(n)}(\mathbf{y})$ — непересекающиеся, всего их имеется $\varkappa=\varkappa_1\cdot\ldots\cdot\varkappa_m$ и $S^{(n)}=\bigcup_{\mathbf{y}\in Y} S^{(n)}(\mathbf{y})$. В каждом подмножестве $S^{(n)}(\mathbf{y}_i)$ $(i=\overline{1,\varkappa})$ имеется по одному глубокому состоянию, которые теперь будем обозначать $\bar{s}_1,\ldots,\bar{s}_\varkappa$. Из условия (3.1) и конструкции автомата $A_\varepsilon^{(n)}$ следует регулярность цепи M_n, поэтому существуют финальные вероятности состояний $\pi(s,n)$ для всех $s\in S^{(n)}$. Для каждого управления $\mathbf{y}\in Y$

$$\pi(\mathbf{y},n)=\sum_{s\in S^{(n)}(\mathbf{y})}\pi(s,n).$$

Из регулярности цепи M_n следует, что

$$\lim_{t\to\infty} W_i(t,A_\varepsilon^{(n)})=\sum_{\mathbf{y}\in Y}[q_i(\mathbf{y})\cdot 1+p_i(\mathbf{y})\cdot 0]\cdot\pi(\mathbf{y},n)=\sum_{\mathbf{y}\in Y}\rho_i W_i(\mathbf{y})\,\pi(\mathbf{y},n). \tag{4.1}$$

Поэтому условие (1.4) означает

$$\min_{1\leqslant i\leqslant k}\sum_{\mathbf{y}\in Y}\rho_i W_i(\mathbf{y})\,\pi(\mathbf{y},n)>\min_{1\leqslant i\leqslant k}\rho_i W_i^0-\varepsilon. \tag{4.2}$$

Тогда, если мы покажем, что последовательность автоматов $\{A_\varepsilon^{(n)}\}$ такова, что в цепи M_n

$$\lim_{n\to\infty} \sum_{y\in Y_c} \pi(y, n) = 1, \tag{4.3}$$

то для данного $\varepsilon>0$ мы найдем автомат $A^{(n_\varepsilon)}$ из этой последовательности, удовлетворяющий (1.4). Действительно, из неравенства

$$\min_{1\leqslant i\leqslant k} \sum_{y\in Y} \rho_i W_i(y)\pi(y, n) \geqslant \sum_{y\in Y} \min_{1\leqslant i\leqslant k} \rho_i W_i(y)\pi(y, n)$$

и определения множества слейтеровских управлений Y_c следует, что при выполнении (4.3) для данного $\varepsilon>0$ найдется n_ε такое, что выполнено (4.2), а значит и (1.4). Таким образом, условие (4.3) является достаточным для (1.4). В этом случае мы можем сказать, что последовательность автоматов $\{A_\varepsilon^{(n)}\}$ ε-оптимальна. Перейдем к доказательству (4.3).

Рассмотрим сужение цепи M_n на множество глубоких состояний так же, как это делается при анализе игры автоматов [4, 6]. Обозначим это сужение M_n', а предельную вероятность глубокого состояния $\bar{s}_j$ $(j=\overline{1, \varkappa})$ в цепи M_n' обозначим $\pi'(\bar{s}_j)$.

Лемма 1. Существуют положительные числа c_1 и c_2 такие, что для всех $n\geqslant 1$ и $j=\overline{1, \varkappa}$ справедливы неравенства

$$c_1\pi'(\bar{s}_j) \leqslant \pi(y_j, n) \leqslant c_2\pi'(\bar{s}_j).$$

Доказательство. Пусть $S_0^{(n)}(j)$ — такое подмножество $S^{(n)}(y_j)$, что всякое $s\in S_0^{(n)}(j)$ соответствует случаю, когда хотя бы один автомат памяти находится в состоянии глубины 0. В множестве $S_0^{(n)}(j)$ содержится $\alpha=(n+1)^k-n^k$ элементов. Из соотношения (3.4) следует, что

$$\pi(\bar{s}_j, n) \geqslant \sum_{s\in S_0^{(n)}(j)} \pi(s, n). \tag{4.4}$$

Обозначим $\tilde{S}^{(n)}(j) = S^{(n)}(y_j) \setminus S_0^{(n)}(j)$; $\tilde{\pi}(y_j, n) = \sum_{s\in\tilde{S}^{(n)}(j)} \pi(s, n)$. Рассуждая полностью аналогично тому, как это делается при анализе игры автоматов, можно показать, что найдется число $c_j>1$ такое, что

$$\pi(\bar{s}_j, n) \leqslant \tilde{\pi}(y_j, n) \leqslant c_j\pi(\bar{s}_j, n). \tag{4.5}$$

Из (4.4), (4.5) следует $\pi(\bar{s}_j, n) \leqslant \pi(y_j, n) \leqslant (c_j+1)\pi(\bar{s}_j, n)$, а поэтому

$$\frac{1}{c} \leqslant \sum_{j=1}^{\varkappa} \pi(\bar{s}_j, n) \leqslant 1,$$

где $c = 1 + \max_{1\leqslant j\leqslant\varkappa} c_j$. Поскольку финальные вероятности $\pi'(\bar{s}_j)$ и $\pi(\bar{s}_j, n)$ связаны, согласно [19], соотношением

$$\pi'(\bar{s}_j) = \pi(\bar{s}_j, n)\left[\sum_{j=1}^{\varkappa} \pi(\bar{s}_j, n)\right]^{-1},$$

то, используя полученные неравенства, находим, что

$$\frac{1}{c}\pi'(\bar{s}_j) \leqslant \pi(y_j, n) \leqslant c\pi'(\bar{s}_j).$$

Осталось положить $c_1 = \frac{1}{c}$, $c_2 = c$.

Мы сможем определить асимптотические значения финальных вероятностей всех управлений $y_j\in Y$ при $n\to\infty$ для исходной цепи M_n, если сумеем

оценить переходные вероятности p_{ij}' между глубокими состояниями $\bar{s}_i$ и $\bar{s}_j$, т. е. между состояниями цепи M_n'. Оказывается, что для алгоритма управления $A_\varepsilon^{(n)}$ удается получить более точные оценки, чем в случае управления с помощью прямого произведения «доверчивых» автоматов.

Лемма 2. Существуют положительные числа c_3 и c_4 такие, что при всех n, i, j $(i, j=\overline{1, \varkappa};\ i\neq j)$

$$c_3 [\max_{1\leqslant l\leqslant k} p_l(\mathrm{y}_i)]^n \leqslant p_{ij}' \leqslant c_4 [\max_{1\leqslant l\leqslant k} p_l(\mathrm{y}_i)]^n.$$

Доказательство. Правое неравенство доказывается точно так же, как и при анализе игры «доверчивых» автоматов Роббинса — Кринского [4, 6]. Докажем левое неравенство. Пусть автомат памяти $D_h^{(n)}$ таков, что $\max\limits_{\leqslant l\leqslant k} p_l(\mathrm{y}_i)$ достигается на $p_h(\mathrm{y}_i)$ и пусть состояние цепи $s'\in S^{(n)}(\mathrm{y}_i)$ соответствует глубине 1 автомата $D_h^{(n)}$ и глубине n остальных автоматов. Тогда, согласно (3.3) и (3.4)

$$p_{ij}' \geqslant \frac{1}{\varkappa} p_h(\mathrm{y}_i)\, p\langle \bar{s}_i \,|\, s'\rangle,$$

поскольку из состояния s' с вероятностью $\frac{1}{\varkappa} p_h(\mathrm{y}_i)$ за два шага произойдет переход на новое управление y_j, а точнее — переход в глубокое состояние $\bar{s}_j$. Оценим вероятность перехода $p\langle \bar{s}_i | s'\rangle$ в цепи M_n из глубокого состояния $\bar{s}_i$ в состояние s'. Пусть состоянию цепи $s\in S^{(n)}(\mathrm{y}_i)$ отвечает ситуация, когда автоматы $D_l^{(n)}$ $(l=\overline{1, k})$ находятся в состояниях глубины λ_l. Тогда, рассматривая k независимых случайных блужданий, совершаемых каждым из автоматов памяти $D_l^{(n)}$, можно установить, что в цепи M_n переход из $\bar{s}_i$ в s происходит за $n-\min\limits_{1\leqslant l\leqslant k}\lambda_l$ тактов, а вероятность этого перехода равна

$$\prod_{l=1}^{k} q_l p_l^{n-\lambda_l}.$$

Отсюда следует, что

$$p\langle \bar{s}_i \,|\, s'\rangle = [p_h(\mathrm{y}_i)]^{n-1} \prod_{l=1}^{k} q_l(\mathrm{y}_i).$$

Используя предыдущее неравенство, получим необходимую оценку снизу для вероятности p_{ij}'.

Леммы 1 и 2 позволяют доказать нужное нам основное утверждение *.

Теорема. Сумма финальных вероятностей слейтеровских управлений стремится к 1 при росте глубины памяти.

Доказательство. Напомним, что множество слейтеровских управлений Y_c — множество управлений, на которых достигается

$$\max_{\mathrm{y}\in Y} \min_{1\leqslant l\leqslant k} \rho_l W_l(\mathrm{y}).$$

Поскольку вероятность штрафа $p_l(\mathrm{y}_i)=1-\rho_l W_l(\mathrm{y}_i)$, то достаточно показать, что для любых управлений y_i и y_j, если $\max\limits_l p_l(\mathrm{y}_i) > \max\limits_l p_l(\mathrm{y}_j)$, то

$$\lim_{n\to\infty} \frac{\pi(\mathrm{y}_i, n)}{\pi(\mathrm{y}_j, n)} = 0. \tag{4.6}$$

Из очевидного неравенства $\pi'(\bar{s}_j) \geqslant \pi'(\bar{s}_i) p_{ij}' + \pi'(\bar{s}_j) p_{jj}'$ следует, что $\pi'(\bar{s}_j)/\pi'(\bar{s}_i) \geqslant p_{ij}'/(1-p_{jj}')$. Из леммы 2 следует, что

$$p_{jj}' = 1 - \sum_{f\neq j} p_{jf}' \geqslant 1 - (\varkappa - 1)\, c_4 [\max_l p_l(\mathrm{y}_j)]^n,$$

*) Внимание автора на это утверждение обратил Е. Т. Гурвич.

поэтому

$$\frac{1}{1-p_{jj'}} \geqslant \frac{1}{(\varkappa-1)\,c_4\,[\max\limits_l p_l(\mathrm{y}_j)]^n}.$$

Пользуясь этим неравенством и левым неравенством из леммы 2, получаем, что

$$\frac{\pi'(\bar{s}_j)}{\pi'(\bar{s}_i)} \geqslant c\,\frac{[\max\limits_l p_l(\mathrm{y}_i)]^n}{[\max\limits_l p_l(\mathrm{y}_j)]^n},$$

где c — положительное число, откуда следует

$$0 \leqslant \lim_{n\to\infty} \frac{\pi'(\bar{s}_i)}{\pi'(\bar{s}_j)} \leqslant \frac{1}{c} \lim_{n\to\infty} \left[\frac{\max\limits_l p_l(\mathrm{y}_j)}{\max\limits_l p_l(\mathrm{y}_i)}\right]^n = 0.$$

Пользуясь леммой 1, получаем соотношение (4.6). Теорема доказана.

5. Заключение. Обучающийся автомат с постоянной структурой, который взаимодействует не с обычной стационарной случайной средой, а с несколькими «учителями», впервые рассматривался в [20]. В этой работе предполагалось наличие априорно корректного поведения, которому автомат должен научиться с помощью k учителей, и рассматривался случай $m=1$. Формально это предположение означает, что существует единственное действие y_γ такое, что для любого другого действия y_j $(j=\overline{1,\varkappa_1})$

$$W_i(y_\gamma) > W_i(y_j) \tag{5.1}$$

для всех $i=\overline{1,k}$, т. е. в множестве слейтеровских управлений Y_c содержится единственный элемент y_γ. В [21] рассматривался стохастический автомат с переменной структурой, взаимодействующий с такой средой из k учителей, в которой существует действие y_β такое, что

$$\sum_{i=1}^{k} W_i(y_\beta) > \sum_{i=1}^{k} W_i(y_j) \quad \text{для всех } j=\overline{1,\varkappa_1}\ (j\neq\beta). \tag{5.2}$$

Предложенный подход, в отличие от [20] и [21], не требует допущений типа (5.1) или (5.2) и позволяет рассматривать поведение автомата в среде из нескольких учителей с точки зрения теории принятия решений при многих критериях.

ЛИТЕРАТУРА

1. *Цетлин М. Л.* Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969.
2. *Варшавский В. И.* Коллективное поведение автоматов. М.: Наука, 1973.
3. *Гурвич Е. Т.* Об асимптотическом поведении автоматов в играх.— В сб.: Тр. Международного симпозиума ИФАК «Дискретные системы», т. 4. Рига: Зинатне, 1974.
4. *Гурвич Е. Т.* Метод асимптотического исследования игр автоматов.— Автоматика и телемеханика, 1975, № 2.
5. *Гурвич Е. Т.* Простейшая модель выбора действия в условиях неопределенности.— Автоматика и телемеханика, 1976, № 11.
6. *Срагович В. Г.* Теория адаптивных систем. М.: Наука, 1976.
7. *Назин А. В., Позняк А. С.* Адаптивный выбор вариантов. М.: Наука, 1986.
8. *Lakshmivarahan S.* Learning algorithms: theory and applications. N. Y.: Springer, 1981.
9. *Wheeler R. M., Narendra K. S.* Learning models for decentralized decision making.— Automatica, 1985, v. 21, N 4.
10. *Назин А. В.* Игровая задача адаптивного выбора вариантов и алгоритм ее решения.— Автоматика и телемеханика, 1983, № 11.
11. *Гермейер Ю. Б.* Введение в теорию исследования операций. М.: Наука, 1971.
12. *Подиновский В. В., Ногин В. Д.* Парето-оптимальные решения многокритериальных задач. М.: Наука, 1982.

13. *Гермейер Ю. Б., Ватель И. А.* Игры с иерархическим вектором интересов.— Изв. АН СССР. Техн. кибернет., 1974, № 3.
14. *Варшавский В. И., Поспелов Д. А.* Оркестр играет без дирижера. М.: Наука, 1984.
15. *Наппельбаум Э. Л., Поспелов Д. А.* Субъективное структурирование задачи в случае коллективного принятия решения.— В кн.: Нормативные и дескриптивные модели принятия решений. М.: Наука, 1981.
16. *Поспелов Д. А.* Введение в теорию вычислительных систем. М.: Сов. радио, 1972.
17. *Robbins H.* A sequential decision problem with finite memory.— Proc. of Nat. Acad. of Science of USA, 1956, v. 42, N 3.
18. *Bacon G. C.* The decomposition of stochastic automata.— Information and Control, 1964, v. 7, N 3.
19. *Кемени Дж., Снелл Дж.* Конечные цепи Маркова. М.: Наука, 1970.
20. *Koditschek D. E., Narendra K. S.* Fixed structure automata in a mutliteacher environment.— IEEE Trans. on Systems, Man, and Cybernetics, v. SMC – 7, 1977, N 8.
21. *Baba N.* The absolutely expedient nonlinear reinforcement schemes under the unknown multiteacher environment.— IEEE Trans. on Systems, Man, and Cybernetics, 1983, v. SMC – 13, N 1.

Москва

Поступила в редакцию
19.VIII.1986