

В. Г. СРАГОВИЧ

ТЕОРИЯ
АДАПТИВНЫХ
СИСТЕМ



ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
Москва 1976

518

С 75

УДК 519.95

Теория адаптивных систем. В. Г. Срагович. Главная редакция физико-математической литературы изд-ва «Наука», М., 1976.

Теория адаптивных управляющих систем изучает способы управления в условиях неопределенности, когда сведения об объекте недостаточны. Системы должны обеспечить достижение цели любым объектом некоторого класса. Такая постановка задачи стимулируется многими приложениями.

Книга посвящена математической теории адаптивных систем управления. Главное внимание уделено конструкциям и свойствам конкретных адаптивных систем для распространенных классов объектов управления. Сначала рассматриваются системы управления объектами из простейших классов, затем — из все более сложных.

Книга предназначается студентам, аспирантам и научным работникам в области прикладной математики.

Книга содержит 11 илл., библ. 93 назв.

ОГЛАВЛЕНИЕ

Предисловие	5
Введение	7
Г л а в а I. Управляемые случайные процессы	16
§ 1. Объекты и системы управления	16
§ 2. Понятие случайного процесса	18
§ 3. Управляемые случайные процессы	25
§ 4. Цели управления	29
§ 5. Постановка задач в классической теории управления	36
Г л а в а II. Адаптивные системы управления	39
§ 1. Определение адаптивных систем управления	39
§ 2. Обучаемые системы	41
§ 3. Ассоциированные марковские процессы	49
§ 4. Общие замечания об адаптивных системах	52
Г л а в а III. Автоматы для управления однородными процессами с независимыми значениями	61
§ 1. Постановка задачи	61
§ 2. Конечные автоматы для бинарных ОПНЗ	64
§ 3. Конечные автоматы для ОПНЗ	80
§ 4. Автоматные алгоритмы управления ОПНЗ (не бинарный случай)	83
§ 5. δ -оптимальные автоматы	85
§ 6. Примеры δ -автоматов	98
§ 7. Асимптотически оптимальные автоматы	116
Г л а в а IV. Автоматы для неоднородных ПНЗ	125
§ 1. Постановка задачи	125
§ 2. Конечные автоматы для бинарных ПНЗ	126
§ 3. Оценивающие автоматы	129
§ 4. δ -автоматы	139
§ 5. Добавления	141
Г л а в а V. Рекуррентные процедуры управления ОПНЗ	148
§ 1. Постановка задачи	148
§ 2. Задача о выполнении плана	149
§ 3. Задача о максимизации выигрыша	156
§ 4. Приложение рекуррентных процедур к задаче прогноза стационарных последовательностей	160
§ 5. Стохастическое программирование	162

Г л а в а VI. Автоматные методы для векторных ОПНЗ	165
§ 1. Децентрализованное управление. Теоретико-игровая интерпретация	165
§ 2. Игры бинарных автоматов. Аналитические методы	170
§ 3. Игры оценивающих автоматов	183
§ 4. Игры автоматов. Моделирование	188
§ 5. Адаптивные иерархические системы	196
Г л а в а VII. Управление обобщенными процессами с независимыми значениями	202
§ 1. Предварительные замечания	202
§ 2. Обобщенные в узком смысле ПНЗ	206
§ 3. Обобщенные в широком смысле ПНЗ	209
§ 4. Синтез автоматов для управления обобщенными ПНЗ	215
§ 5. Свойство адаптивности автоматов $(CD)_{k,n}^{(l,r)}$	218
Г л а в а VIII. Алгоритмы управления марковскими процессами	225
§ 1. Предварительные замечания	225
§ 2. ϵ -оптимальные семейства для эргодических марковских процессов	229
§ 3. Алгоритмы управления конечными марковскими цепями с доходами	239
§ 4. Задачи с целевыми неравенствами	258
Г л а в а IX. Управление процессами общего вида	273
§ 1. Стационарные процессы	273
§ 2. α -однородные процессы	293
Г л а в а X. Об адаптивном управлении процессами с непрерывным временем	301
Примечания	309
Литература	312

ПРЕДИСЛОВИЕ

Предмет этой монографии — математическая теория адаптивных систем управления. Эта теория, представляющая собой часть теории управления, возникла и развивается благодаря все более широкому распространению ситуаций, в которых управление объектом протекает при недостатке сведений об управляемом объекте.

В качестве модели объектов управления в книге приняты управляемые случайные процессы с дискретным временем. Процессам с непрерывным временем посвящена всего одна глава. Среди различных форм управляемых процессов и управляющих систем особое внимание мы уделяем дискретным (и даже конечным) процессам и системам. В частности, много места отведено адаптивным системам, представляющим собой конечные автоматы. Дискретность позволяет достичь строгости изложения, оставляя неприкосновенными математическую значимость и практическую ценность. Последнее связано с тем, что базой автоматических управляющих систем стала цифровая вычислительная техника.

При отборе материала для монографии приняты следующие условия: необходимые — рассматриваемые проблемы, задачи и вопросы должны иметь точную постановку и строгое решение, достаточные — они должны соответствовать научным интересам и вкусам автора. Впрочем, в ряде случаев пришлось еще учесть ограничения на объем книги.

В книге, по-видимому, удалось отразить все основные идеи и принципы адаптивного управления. К сожалению, ее объем не позволил сообщить хотя бы часть конкретных результатов, основанных на каждой из таких идей или принципов. Однако вошедшие в текст факты изложены подробно, за исключением относящихся к методу стохастической аппроксимации. Ему посвящено много солидных

монографий, которые дублировать нецелесообразно. Поэтому в главе о стохастической аппроксимации содержится лишь сводка основных результатов. Нельзя вовсе опустить этот метод без искажения общей картины теории адаптивных систем.

В конце книги, в примечаниях, собраны ссылки на литературу. Библиография содержит преимущественно источники, использованные в основном тексте. В некоторых из них приведены списки многих других работ.

Следующим проблемам, вероятно, следовало уделить внимание, но по разным причинам их пришлось опустить. Прежде всего, это адаптивные подходы к распознаванию и идентификации. Далее, способы использования адаптивных систем для управления реальными объектами. Наконец, полностью опущена история развития теории, видное место в которой занимают отечественные ученые.

Предполагается, что читатель знаком с проблематикой теории управления и владеет университетским курсом теории вероятностей.

Автор сердечно благодарит В. Н. Фомина и Г. А. Агасандяна за дискуссии и многочисленные критические замечания, способствовавшие улучшению рукописи.

В. Г. Срагович

ВВЕДЕНИЕ

Первые задачи зародившейся теории управления состояли в стабилизации характеристик объекта управления, удержании его параметров в заданных пределах. Так возникла теория автоматического регулирования, включившая в себя затем также проблемы устойчивости. Частью этой теории стали методы синтеза регуляторов, обеспечивающих протекание переходного режима заданным образом. Поведение объекта (обычно речь идет о его движении) описывается дифференциальными уравнениями, подразумевается, что известны не только структура уравнений, но и значения всех входящих в них величин.

В дальнейшем автоматическое регулирование охватило стохастические объекты. Для проектирования следящих систем были созданы методы расчета фильтров (для выделения полезного сигнала из помех) и прогнозирующих устройств. Здесь снова предполагают точно и полно известными вероятностные свойства объекта, который описывается понятием «стационарный случайный процесс». От обеспечивающих достижение цели систем управления требуется быть в некотором смысле наилучшими, а именно, получить результаты (оперируя над процессом) с наименьшей среднеквадратической ошибкой. Этот подход нуждается в предварительном полном описании объекта управления (в рамках требований соответствующих методов). Не могло быть так, чтобы какие-то функции, параметры и т. п., фигурирующие в математической модели объекта, оставались неизвестными, неопределенными. Это обстоятельство будет более ясным, если вспомнить, что методы синтеза систем управления создавались преимущественно для реализации на ЦВМ, оперирующих только с числами, а не с формулами.

Существенной чертой охарактеризованных здесь разделов теории управления является требование полноты описания объектов управления. Поэтому в случае пробелов в знаниях о каком-либо объекте приходится заранее, перед тем как начать синтез системы управления, изучить объект. Допускается, что система управления измеряет текущее состояние объекта и пользуется поступающими по каналу обратной связи сведениями для осуществления наперед указанной цели. Однако не имеется в виду уточнение свойств объекта и соответствующая перестройка управляющей системы. Если, например, «оптимальный» фильтр, выделяющий из помехи полезный сигнал, найден по неправильно заданной корреляционной функции, то сколь угодно длительное его функционирование никогда не станет желаемым, т. е. отыскивающим сигнал в помехе с минимальной ошибкой. Теория управления не предусматривает коррекцию самой системы управления по мере уточнения свойств объекта. На практике такую реконструкцию осуществляют после настоящий эксплуатационников, установивших экспериментально (или интуитивно) несоответствие системы своему назначению.

Условимся для краткости называть классической теорией управления те ее разделы, которые основываются на постулате, что известна точная и полная математическая модель объекта. За долгие годы существования она не только стала заметной и важной областью науки, но и успешно зарекомендовала себя в многочисленных приложениях.

В 50-х годах стала назревать необходимость в расширении сферы действия теории управления, в частности, в отказе от полного описания объектов. Причиною тому служат не столько естественность обобщений теории, сколько запросы промышленности, связи, экономики и иных сфер деятельности. Их развитие сделало необходимым управление такими объектами, для которых не только отсутствует адекватная математическая модель, но иногда даже самые общие, качественные закономерности изучены недостаточно. Приведем некоторые примеры технологических процессов, для которых неизвестна кинетика. В первую очередь сюда относится почти вся химическая промышленность и в первую очередь нефтехимия. Ука-

жем, например, каталитический крекинг нефти, пиролиз бензина (для производства этилена). Кроме того, можно назвать производства парафинов, антибиотиков и многое, многое другое. Среди иных отраслей промышленности следует отметить металлургию, для которой в немногих случаях известны строгие количественные законы протекания технологических процессов. Далее, к числу процессов без точного описания относится трубопрокатное производство и даже многие сборочные линии. Этот очень краткий список дает представление о степени распространенности объектов, для которых синтез управляющих систем на базе классической теории управления затруднителен или невозможен. Неудивительно поэтому, что столь дружный запрос ключевых промышленных направлений привел к быстрому прогрессу науки.

Зародившийся подход к новым «неклассическим» задачам управления основывается на идее приспособления управляющей системы к свойствам конкретного объекта, о котором заранее известно всего лишь, к какому классу управляемых объектов он относится.

В ходе функционирования системы управления происходит ее приспособление к находящемуся перед ней объекту, которое может проявляться, например, в смене параметров и (или) структуры системы для того, чтобы она по прошествии некоторого, по возможности короткого, промежутка времени могла обеспечивать назначенную цель управления.

Напрашиваются биолого-физиологические аналогии: живые существа обладают способностью приспабливаться, адаптироваться к меняющимся в довольно широких пределах условиям обитания. Это относится не только к целостному организму, но и к отдельным его органам и функциям. Становление новой проблематики управления совпало по времени с всплеском увлечения биологией, наложившим отпечаток на терминологию и вызвавшим надежду, что скорое открытие всех загадок живого позволит строить совершенные технические системы управления.

Один из первых принципов синтеза нового типа систем (приспосабливающихся, или, как мы будем говорить, адаптивных) заключался в объединении двух подходов:

изучение природы объекта и применение методов классической теории управления — вычислении приводящего к цели управления по уже установленным (вообще говоря, с ошибками) характеристикам объекта. Такой метод делает теорию адаптивных управляющих систем расположенной «между» классической теорией управления и математической статистикой или, более общо, теорией идентификации. Он означает сохранение предположения, что для успешного управления необходимо знать по возможности больше об объекте. Этот «идентификационный» подход был плодотворен на ранних этапах становления адаптивных систем. Не утерял он своего значения и поныне (в тех нечастых случаях, когда он применим). Однако постепенно было достигнуто понимание того, что можно успешно управлять, не утруждая себя детальным изучением свойств объекта. На справедливость этого наталкивают даже физиологические соображения: адаптируясь к внешним условиям, организм не утруждает себя исследованиями по метеорологии, физике и химии атмосферы и т. п. Помимо этих косвенных соображений (нелишне заметить, что такого sorta «аргументы» часто считались доказательными на ранних этапах развития представлений об адаптивных системах), есть более веские доводы об ограниченной эффективности «идентификационного» подхода. Один из них состоит в том, что фактическое отыскание оптимального управления обычно бывает весьма трудоемким, более того, часто отсутствуют методы его вычисления. Другой довод заключается в существовании таких объектов, подчеркнем — достаточно простого вида, для которых идентификация невозможна. Рассмотрим один пример.

Задан класс объектов, описываемых разностными уравнениями вида

$$x_{t+1} = ax_t + by_t + \zeta_t,$$

где x_t — числовая характеристика объекта в момент t , y_t — управление, ζ_t — внешнее возмущение, a и b — постоянные коэффициенты. Класс характеризуется всевозможными тройками $\{a, b, (\zeta_t)\}$ и при управлении

a, b, ζ_t неизвестны, а ζ_t не наблюдаемо. Ясно, что из последовательных соотношений

$$\begin{aligned}x_2 &= ax_1 + by_1 + \zeta_1, \\x_3 &= ax_2 + by_2 + \zeta_2, \\x_4 &= ax_3 + by_3 + \zeta_3, \\&\dots\end{aligned}$$

коэффициенты управления, и даже какой-нибудь один, определить нельзя. Поэтому, какая бы цель ни стояла перед системой управления, последняя не может рассчитывать на большую информацию, чем на сведения о структуре уравнения, т. е. об его линейности. Синтез управления, которое обеспечивало бы, например, выполнение с некоторого момента времени неравенства $|x_t| < r$, должно базироваться на каких-то новых идеях. Развитие таких идей, принципов, методов и составило предмет теории адаптивных систем. Некоторые из них уже не только оказались плодотворными в теоретических исследованиях, но и с лучшей стороны зарекомендовали себя в реальных системах управления.

В настоящее время теория адаптивных управляющих систем вышла из начальной стадии развития и уже является полноправным разделом теории управления. Помимо научного значения, она стала основой построения многих действующих систем. Есть основания считать, что она в перспективе окажется центром теории управления. В самом деле, технический прогресс дает все больше и больше таких производств, технологических процессов, управлять которыми необходимо, но сведений о которых нет и ожидать их появления нереально.

На пути развития теории адаптивных систем долгое время стояло одно препятствие. Оно заключалось в разобщенности разных групп исследователей, сказывающейся не только в разных задачах и методах их решения (это, вероятно, даже положительный факт для ускоренного развития), но и в разной терминологии, препятствующей взаимному пониманию. В потоке «адаптивных» работ оказалось много таких, в которых слово «адаптация» фигурирует без достаточных оснований. Одной причиной тому служит присущая инженерным и прикладным работам

(а они долгое время составляли подавляющую долю публикаций в этой области) беззаботность относительно определений*) и вообще четкой системы основных понятий. Тем не менее можно установить и сформулировать, что именно большинство исследователей понимает под словами «адаптивная система». Пусть задан класс объектов, относительно желаемых свойств которых сформулирована цель управления. Адаптивная система — это такая управляющая система, которая в ходе управления любым объектом класса за конечное время достигает цель. В отличие от управляющих систем классической теории, адаптивные системы приводят к цели лишь по прошествии некоторого времени после начала функционирования. Это — неизбежное последствие неопределенности управляемого объекта и необходимости системе управления «привыкнуть», «приспособиться» к нему. Трудно и, как правило, невозможно заранее указать необходимое время «обучения». Поэтому задачи адаптивного управления ставятся на неограниченном интервале времени.

Перейдем к изложению способов задания и классификации объектов управления. Время t примем дискретным ($t = 0, 1, \dots$). Прежде всего, необходимо выбрать два пространства — фазовое $X = \{x\}$ и управлений $Y = \{y\}$. Тогда эволюцию объекта можно записать с помощью уравнения

$$x_{t+1} = g(x_t, \dots, x_1; y_t, \dots, y_1, y_0), \quad t \geq 0.$$

В более простых случаях уравнение имеет такой вид:

$$x_{t+1} = g(x_t, \dots, x_{t-l+1}; y_t, \dots, y_{t-l}), \quad t \geq l,$$

к которому следует добавить начальные условия. Удобная классификация объектов основывается на «порядках» l описывающих их уравнений. Простейшие объекты характеризуются зависимостями

$$x_{t+1} = g(y_t, \dots, y_{t-l}), \quad t \geq l,$$

*) Вот один из характерных примеров: «Термин самонастраиваящаяся (adaptive) система применяют для широкого класса систем, куда входят и некоторые системы с обратной связью, в работе которых проявляется действие, напоминающее приспособление в организме или обществе» (Р. Драйек, Тр. Международного симпозиума (ИФАК), М., «Наука», 1964, стр. 37).

в которых состояние в каждый момент зависит лишь от приложенных до того управлений, но не от предшествующих состояний объекта. Несмотря на кажущуюся искусственность, такие объекты долгое время оставались главным предметом исследования теории адаптивных процессов.

Следующий класс объектов подчиняется уравнениям первого порядка

$$x_{t+1} = g(x_t; y_t, \dots, y_{t-1})$$

с начальным условием. Дальнейшее усложнение очевидно и заключается в повышении порядка уравнения.

Приведенная классификация недостаточна потому, что не охватывает объекты, в функционирование которых вмешивается случайность. Поэтому следует расширить математическое описание объекта. Тогда можно было бы строить единообразную теорию, включающую в себя и детерминированные, и стохастические объекты. Это можно сделать, если воспользоваться понятием управляемого случайного процесса. Оно означает, что в пространстве X задано семейство случайных процессов ξ_t (где либо $t = 0, 1, 2, \dots$, либо $t = \dots, -1, 0, 1, \dots$) с распределениями вероятностей, зависящими от управлений $y \in Y$. Выделение из этого класса того или другого конкретного процесса производится указанием способа выбора управлений в каждый момент времени. Такие способы, называемые стратегиями, в общем случае задают условными распределениями на пространстве управлений Y , зависящими от истории процесса, — принятых им ранее значений и приложенных управлений. Простейшая форма стратегии является программой, которая заранее на каждый момент времени назначает управление.

Используя термины «стратегия» и «управляемый случайный процесс», можно сказать, что предметом классической теории управления является синтез стратегий, которые приводят к желаемой цели конкретный управляемый случайный процесс. В отличие от нее, теория адаптивных систем изучает стратегии, приводящие к цели класс управляемых случайных процессов.

Дадим теперь обзор содержания книги. Математическому описанию объектов управления посвящена гл. I, где даны формальные определения управляемых случай-

ных процессов и стратегий, указаны примеры. В гл. II сформулированы функциональное и конструктивное определения адаптивных управляющих систем и обсуждаются относящиеся к ним общие вопросы. Важную роль играет установленная здесь представимость всякой адаптивной системы в виде автомата, хотя в большинстве случаев и бесконечного. Состояниями автомата служат правила выбора управлений, и функционирование адаптивной системы представляет собой целенаправленное случайное служдание по множеству правил.

Гл. III—VI посвящены адаптивным системам, управляющим простейшими управляемыми случайными процессами, теми, которые в детерминированном варианте изображаются уравнениями $x_t = g(y_{t-1})$. Средствами достижения выдвигаемых целей управления служат автоматы (с конечным и счетным множествами состояний) и рекуррентные процедуры, получившие широкую известность под названием метода стохастической аппроксимации. В гл. VII рассмотрены системы управления процессами более общего типа, подчиняющиеся уравнениям $x_t = g(y_{t-1}, \dots, y_{t-l-1})$ при $l \geq 2$.

Методы адаптивного управления марковскими процессами излагаются в гл. VIII. Эти процессы в детерминированном случае подчинены уравнениям типа $x_{t+1} = g(x_t, x_{t-1}, \dots, x_{t-l}; y_t)$ и представляют собой управляемые динамические системы, преимущественным образом встречающиеся в практических задачах. Выдвигаемые для них типичные цели управления преимущественно образуют две группы: оптимизационные задачи и задачи устойчивости. Мы рассматриваем адаптивные системы для каждой из этих групп целей.

Процессы общего вида мы определяем как подчиняющиеся уравнениям

$$x_{t+1} = g_t(x_t, \dots, x_1; y_t, \dots, y_1, y_0),$$

т. е. в каждый момент времени они зависят от всей предыстории. В гл. IX построены и исследованы адаптивные системы для некоторых классов процессов общего вида. Роль таких процессов в приложениях пока не слишком велика, но синтез соответствующих адаптивных систем

важен по той причине, что выясняются возможные границы теории адаптивных систем управлений.

Перечисленные проблемы и задачи относятся к управляемым случайнм процессам, которые протекают в дискретном времени. Здесь уместно напомнить, что первые исследователи по управляющим адаптивным системам занимались процессами с непрерывным временем.

В работах Б. Н. Петрова, А. А. Красовского, Г. С. Поспелова, а затем и других ученых построены адаптивные управляющие системы для классов объектов, описываемых дифференциальными уравнениями. Изложение достигнутых результатов требует много места. Поэтому мы вынуждены ограничить полноту рассмотрения этой темы. В гл. X описаны лишь два типа адаптивных систем, функционирующих в непрерывном времени.

ГЛАВА I

УПРАВЛЯЕМЫЕ СЛУЧАЙНЫЕ ПРОЦЕССЫ

§ 1. Объекты и системы управления

Буква t всюду означает параметр — дискретное неограниченное время. Далее рассматриваются два возможных варианта: 1) существует начальный момент t_0 , без ограничения общности можно принять $t_0 = 1$, тогда время t пробегает значения $1, 2, \dots$; 2) начального момента нет и время принимает все целые значения, т. е. $t = \dots, -1, 0, 1, \dots$.

Пусть дана последовательность x_t элементов множества $X = \{x\}$. Мы используем такие обозначения: x_s^t , $t \geq s$, есть $(t-s)$ -мерный вектор $x_s^t = (x_{s+1}, \dots, x_t)$, x^t — либо конечномерный вектор (x_1, \dots, x_t) , если существует начальный момент времени, либо бесконечномерный (\dots, x_{t-1}, x_t) в противном случае.

Множество X служит впредь фазовым пространством эволюционирующего во времени объекта. Правило эволюции в общем виде задается равенством

$$x_{t+1} = g(x^t),$$

быть может, с начальным условием $(x_1 = x)$. В распространенных случаях оно имеет специальную форму

$$x_{t+1} = g(x_{t-l+1}^t),$$

где $l \geq 1$ — постоянное целое число, называемое «порядком» объекта. Здесь приходится задавать l начальных значений x_1, \dots, x_l , по которым определяют x_{l+1} , а за ним и дальнейшие члены последовательности x_t .

Допустим, что объект управляем. Это означает, что задано пространство управлений $Y = \{y\}$, элементы которого явно фигурируют в эволюционном уравнении

$$x_{t+1} = g(x^t, y^t).$$

Управляемый объект функционирует так: в начальный момент $t = 0$ выбирается управление y_0 , оно определяет следующее значение процесса $x_1 = g(y_0)$. В момент $t = 1$ задают y_1 и получают $x_2 = g(x_1; y_0, y_1)$ и т. д. Будем

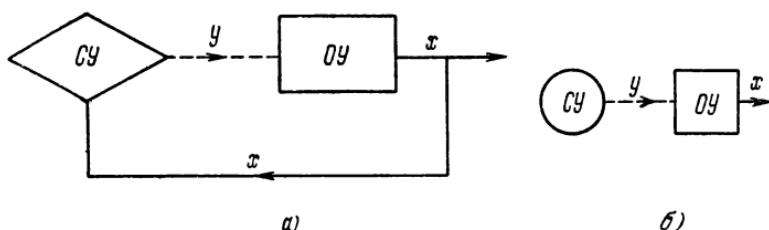


Рис. 1.

считать, что выбор значений управления осуществляет система управления.

Связь системы управления (СУ) и объекта управления (ОУ) имеет одну из двух форм: либо с каналом обратной связи, либо без него (разомкнутое управление) (рис. 1, а, б). Слова «взаимодействие объекта управления и системы управления» означают, что значения процесса x_i «измеряются» и поступают на вход системы управления, которая по ним находит очередные управления и отправляет их объекту. Правило отыскания значений управления подчиняется условию, чтобы достигалась некоторая заданная цель, т. е. траектория x_t в фазовом пространстве обладала предписанным свойством. Если объект обозначить буквой O , а систему управления S , то для составного объекта, получающегося при взаимодействии O и S , примем обозначение $S \otimes O$.

До этого момента подразумевалось, что и объект, и система управления являются детерминированными и во многих случаях это разумно. Но не всегда. Часто объект управления подвержен случайным воздействиям, каналы обратной связи и передачи управляющих сигналов бывают

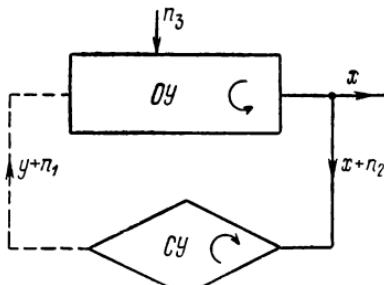


Рис. 2.

засорены шумами. Сверх того, необходимо учитывать, что существуют объекты стохастической природы, среди них отметим системы массового обслуживания, некоторые химические производства и т. д. Деятельность системы управления может быть подвержена ошибкам, случайным сбоям. Схематически сказанное здесь изображено на рис. 2 (n_1, n_2, n_3 означают помехи, шумы, а изогнутые стрелки — наличие «внутренних случайностей»).

В качестве единообразной математической модели объекта управления мы избираем «управляемые случайные процессы», которые включают как частный случай детерминированные объекты. Последние появляются, когда распределения вероятностей вырождены.

§ 2. Понятие случайного процесса

Заданы вероятностное пространство элементарных событий (Ω, \mathcal{F}, P) (где \mathcal{F} — σ -алгебра измеримых множеств из Ω , P — определенная на ней вероятностная мера) и измеримое фазовое пространство (X, \mathfrak{M}) (где \mathfrak{M} — σ -алгебра измеримых множеств из X).

Случайной величиной $\xi(\omega)$ называется измеримое отображение *) $\xi: \Omega \rightarrow X$.

Случайной величине $\xi(\omega)$ со значениями в X отвечает мера на пространстве X , определяемая для любого множества $M \in \mathfrak{M}$ равенством $P(M) = P(\omega: \xi(\omega) \in M)$.

Случайным вектором $\xi(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$ называется измеримое отображение $\xi: \Omega \rightarrow X^n$. Такому вектору соответствует n -мерное распределение

$$P(M_1 \times \dots \times M_n) = P(\omega: \xi(\omega) \in M_1 \times \dots \times M_n)$$

при любых $M_i \in \mathfrak{M}$, $i = 1, \dots, n$.

Случайные величины $\xi_1(\omega), \dots, \xi_n(\omega)$ независимы, если при любых измеримых M_1, \dots, M_n справедливо равенство $P(M_1 \times \dots \times M_n) = \prod_1^n P_i(M_i)$, где P_i — мера на X , отвечающая случайной величине $\xi_i(\omega)$.

* Это означает, что прообразы всех множеств из σ -алгебры \mathfrak{M} принадлежат σ -алгебре \mathcal{F} .

Скажем, что величины из семейства $\{\xi_t(\omega)\}$ независимы, если для любого набора различных индексов (t_1, \dots, t_n) , $n \geq 1$, величины $\xi_{t_1}, \dots, \xi_{t_n}$ независимы.

Случайным процессом называется семейство ξ_t случайных величин.

Процесс ξ_t необрывающийся, если он определен при всех рассматриваемых значениях t .

Случайному процессу сопоставляют меру (распределение вероятностей) на множестве X^∞ всевозможных последовательностей, элементы которых принадлежат X . Выразить ее явным образом удается лишь в исключительных случаях. Обычно оперируют с совокупностью более простых распределений, именуемой системой конечномерных распределений. Последняя определяется следующим образом: для каждого $l \geq 1$ зададим произвольные l моментов времени $t_1 < t_2 < \dots < t_l$ и рассмотрим вероятности

$$\mathbb{P}_{t_1, \dots, t_l}(M_1, \dots, M_l) = \mathbb{P}(\xi_{t_j}(\omega) \in M_j, j = \overline{1, l}).$$

Совокупность их при всех возможных l и t_1, \dots, t_l образует систему конечномерных распределений. Пусть для этой системы выполнены «условия согласованности»:

$$\begin{aligned} 1. \quad \mathbb{P}_{t_{i_1}, t_{i_2}, \dots, t_{i_l}}(M_{i_1}, M_{i_2}, \dots, M_{i_l}) = \\ = \mathbb{P}_{t_1, t_2, \dots, t_l}(M_1, \dots, M_l), \end{aligned}$$

где i_1, i_2, \dots, i_l — произвольная перестановка чисел $1, 2, \dots, l$.

$$\begin{aligned} 2. \quad \mathbb{P}_{t_1, \dots, t_l}(M_1, \dots, M_l) = \\ = \mathbb{P}_{t_1, \dots, t_l, \dots, t_m}(M_1, \dots, M_l, X, \dots, X), \end{aligned}$$

где $l < m$.

Можно доказать, что тогда существует (и притом единственное) распределение на X^∞ , которое на любом конечномерном пространстве (вида X^l) совпадает с исходным соответствующим конечномерным распределением $\mathbb{P}_{t_1, \dots, t_l}$.

В случае дискретного фазового пространства X гарантирована возможность перехода от конечномерных распределений $\mathbb{P}_{t_1, \dots, t_l}$ к системе условных мер (распределений вероятностей)

$$\mu(M | x^l) = \mathbb{P}(\xi_{t_{l+1}} \in M | \xi^l = x^l).$$

Наоборот, от этой последней системы всегда можно перейти к системе конечномерных распределений и, следовательно, задать меру на множестве траекторий X^ω случайного процесса.

Мы будем при задании случайного процесса отдаваться от системы условных распределений вероятностей $\{\mu(\cdot | x^t), t \geq 1\}$. Относительно входящих в нее мер делаем следующие предположения:

1. При каждом $M \in \mathfrak{M}$ функции $\mu(M | x^t)$ измеримы относительно величин, входящих в условие.

2. Построенная по совокупности $\{\mu\}$ система конечномерных распределений удовлетворяет условиям согласованности.

Через φ_t будем обозначать измеримые отображения

$$\varphi : X^t \rightarrow R_1$$

на числовую прямую (с лебеговской мерой).

Рассматриваются измеримые функционалы $\varphi_t = \varphi(\xi^t)$ на траекториях случайного процесса $\xi_t(\omega)$. Ясно, что φ_t — скалярный случайный процесс, для которого можно указать систему условных распределений (или конечномерных распределений), если они известны для исходного процесса. Нас интересуют моменты этого процесса, они заведомо существуют в случае ограниченного функционала. Явная запись математического ожидания $W(t) = E\varphi_t$ такова:

$$W(t) = \int_{X^t} \varphi(x_1, \dots, x_t) \mu(dx_t | x^{t-1}) \mu(dx_{t-1} | x^{t-2}) \dots \\ \dots \mu(dx_2 | x_1) \mu(dx_1).$$

Приведем примеры используемых в дальнейшем случайных процессов.

Последовательность независимых случайных величин характеризуется условием $\mu_{t+1}(M | \xi^t) \equiv \mu_{t+1}(M)$, $M \in \mathfrak{M}$. Если мера μ не зависит явным образом от времени, величины ξ_t одинаково распределены. Функционалы вида $\varphi_t = \varphi(\xi_t)$ на таких процессах представляют собой ска-

лярные процессы того же вида. Их математические ожидания записываются просто:

$$W(t) = \int_X \varphi(x) \mu_t(dx).$$

Изучение последовательностей независимых случайных величин долго оставалось главным предметом теории вероятностей. Мы будем часто использовать

Усиленный закон больших чисел. Пусть ξ_t — последовательность независимых случайных величин. Для выполнения равенства

$$P\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n (\xi_t - E\xi_t) = 0\right) = 1$$

достаточно, чтобы

$$\sum_{t=1}^{\infty} \frac{D\xi_t}{t^2} < \infty.$$

В частном случае одинаково распределенных величин для того, чтобы

$$P\left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n \xi_t = E\xi_1\right) = 1,$$

необходимо и достаточно условие $E|\xi_1| < \infty$.

При выполнении последнего равенства для любого сколь угодно малого $\varepsilon > 0$ найдется такой момент времени τ_ε , что при всех $t > \tau_\varepsilon$ выполняется неравенство

$$\left| \frac{1}{t} \sum_{n=1}^t \xi_n - E\xi_1 \right| < \varepsilon.$$

Этот момент времени случаен, с вероятностью 1 конечен ($P(\tau_\varepsilon = \infty) = 0$) и представляет собой немарковский момент *). Знание зависимости τ_ε от ε позволяет судить

*) Пусть ξ_t — случайный процесс. *Марковским моментом* $\tau(\omega)$ называется случайная величина с целочисленными значе-

о скорости сходимости $\frac{1}{t} \sum_1^t \xi_n$ к пределу $E\xi_1$. Будем считать, что $D\xi_t = 1$.

Обозначим через $G(z)$ χ^2 -распределение с одной степенью свободы. Имеют место следующие факты:

1. Для каждого $z \geq 0$ имеем $\lim_{\epsilon \rightarrow 0} P(\epsilon^2 \tau_\epsilon \leq z) = G(z)$.

2. $E\tau_\epsilon \sim \epsilon^{-2}$.

3. Если существует момент $E\xi_t^{k+1}$, $k = 1, 2, \dots$, то величина τ_ϵ имеет моменты k -го порядка. Если существуют все моменты $E\xi_t^k$, то и τ_ϵ имеет конечные моменты всех порядков.

Марковские процессы характеризуются условием (при всех $M \in \mathfrak{M}$)

$$\mu_{t+1}(M | x^t) \equiv \mu_{t+1}(M | x_t).$$

В правой части этого равенства находится «переходная функция». Процессы, для которых эта функция не зависит явным образом от времени, называют однородными. Функция $\eta_t = h(\xi_t)$ марковского процесса не является, вообще говоря, марковским процессом.

При конечном фазовом пространстве $X = \{x_1, \dots, x_N\}$ будем говорить «марковская цепь» вместо «марковский процесс».

Изложим необходимые для дальнейшего понятия и факты относительно конечных однородных марковских цепей. Их задают посредством квадратных стохастических матриц $\mathcal{P} = \|p_{ij}\|$, где

$$p_{ij} = P(\xi_{t+1} = x_j | \xi_t = x_i)$$

не зависят от времени и удовлетворяют двум условиям

$$p_{ij} \geq 0, \quad \sum_{j=1}^N p_{ij} = 1, \quad i, j = \overline{1, N}.$$

ниями такая, что при каждом t множество $\{\omega: \tau(\omega) < t\}$ измеримо относительно σ -алгебры, порожденной \dots, ξ_{t-1}, ξ_t . Почти наверное конечный марковский момент иногда называют моментом остановки.

Типичным примером марковского момента является (случайный) момент первого попадания процесса ξ_t в заданное множество M фазового пространства.

Вероятности перехода из x_i в x_j за $n \geq 1$ шагов являются элементами n -й степени матрицы \mathcal{P} , т. е.

$$p(\xi_{t+n} = x_j | \xi_t = x_i) = p_{ij}^{(n)},$$

где $\|p_{ij}^{(n)}\| = \mathcal{P}^n$.

В теории марковских процессов (и цепей) значительное место занимают эргодические процессы (и цепи).

Скажем, что состояние x_j следует за x_i , если $p_{ij}^{(n)} > 0$ при некотором n . Состояния x_i и x_j сообщающиеся, если каждое из них следует за другим. Состояние x_i называется *периодическим* (или *циклическим*), если для множества тех n , для которых $p_{ii}^{(n)} > 0$, общий наибольший делитель $d > 1$. Число d — *период состояния* (или *длина цикла*).

Марковская цепь — *эргодическая*, если все ее состояния сообщающиеся и среди них нет периодических. Для эргодичности цепи необходимо и достаточно, чтобы матрица \mathcal{P}^n при некотором n состояла из положительных элементов.

Эргодические марковские цепи обладают следующим важным свойством: на множестве X существует предельное распределение π такое, что

$$\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi(x_j) \geq 0, \quad j = 1, \dots, N,$$

независимо от начального состояния x_i . Скорость приближения к пределу указывает оценка

$$|p_{ij}^{(n)} - \pi(x_j)| < a \cdot e^{-\lambda n}, \quad a > 0, \lambda > 0.$$

Из чисел $\pi(x_j) = \pi_j$ образуем N -мерный стохастический вектор $\pi = (\pi_1, \dots, \pi_N)$. Отыскание предельного распределения является простой в принципе вычислительной задачей решения системы линейных алгебраических уравнений

$$\pi = \pi \mathcal{P}$$

при нормирующем условии $\sum_1^N \pi_i = 1$.

Пусть на множестве состояний эргодической марковской цепи ξ_t задана функция g . Предельное математическое ожидание величины $g(\xi_t)$ по определению равно

$$\mathbf{E}_\pi g = \sum_{j=1}^N g(x_j) \pi_j.$$

Усиленный закон больших чисел для таких цепей выражается равенством

$$\mathsf{P}\left(\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=1}^t g(x_n) = \mathbf{E}_\pi g\right) = 1.$$

Это утверждение переносится на соответствующие классы марковских цепей со счетным множеством состояний.

Марковские цепи с бесконечными множествами состояний появляются различными способами. Один из них состоит в решении стохастических разностных уравнений, имеющих, например, вид

$$\xi_{t+1} = h(\xi_t, \zeta_t),$$

где ζ_t — последовательность независимых случайных величин. Часто встречается линейный случай $\xi_{t+1} = a\xi_t + b\zeta_t$.

Сложные (l -связные) марковские процессы характеризуются условием (при всех $M \in \mathfrak{M}$)

$$\mu_{t+1}(M | x^t) \equiv \mu_{t+1}(M | x_{t-l}^t), \quad l \geq 1.$$

Такие процессы задают, в частности, уравнениями вида $\xi_{t+1} = h(\xi_{t-l}, \zeta_t)$.

Мартингалом (полумартингалом) называется скалярный случайный процесс ξ_t , подчиненный двум условиям:

1. $\mathbf{E} |\xi_t| < \infty, t \geq 1.$

2. С вероятностью 1 выполняется равенство

$$\mathbf{E} (\xi_{t+1} | x^t) = x_t (\mathbf{E} (\xi_{t+1} | x^t) \geq x_t).$$

Следующая теорема является основным результатом теории мартингалов.

Пределная теорема. Если ξ_t — маргингал или полумаргингал и $\sup_t E|\xi_t| < \infty$, то существует случайная величина ξ_∞ , $E|\xi_\infty| < \infty$, такая, что

$$P\left(\lim_{t \rightarrow \infty} \xi_t = \xi_\infty\right) = 1.$$

Если для маргингала или неотрицательного полумаргингала ξ_t при некотором $\alpha > 1$ выполнено условие $\lim_{t \rightarrow \infty} E|\xi_t|^\alpha < \infty$, то существует такая случайная величина ξ_∞ , что

$$\lim_{t \rightarrow \infty} E|\xi_t - \xi_\infty|^\alpha = 0, \quad E|\xi_\infty|^\alpha < \infty.$$

§ 3. Управляемые случайные процессы

Заданы измеримые пространства: фазовое (X, \mathfrak{M}) и пространство управлений (Y, \mathfrak{N}) . Элементы последнего называют управлениями или действиями, через y^t обозначим последовательность y_0, y_1, \dots, y_t (здесь для определенности подразумевается, что t пробегает значения $0, 1, 2, \dots$).

Семейством управляемых условных вероятностей

$$\{\mu_{t+1}(M|x^t, y^t), M \in \mathfrak{M}, t \geq 0\}$$

называется семейство функций, подчиненных условиям:

1. Каждая функция является распределением вероятностей на X при всех последовательностях x^t, y^t .

2. Каждая функция $\mu_{t+1}(\cdot|x^t, y^t)$ измерима по совокупности (x^t, y^t) , т. е. на пространстве $X^t \times Y^{t+1}$.

Мы всегда подразумеваем, что каждое распределение μ_{t+1} существенно зависит по крайней мере от y_t , т. е. фигурирующие в условии y_t, y_{t-1}, \dots не являются все фиктивными переменными. Это исключает из рассмотрения тривиальные случаи.

Пусть на измеримом пространстве (Ω, \mathcal{F}) определены функции $\xi_t(\omega)$, $t \geq 1$, со значениями в X .

Управляемым случайнм процессом называется класс необывающихся случайных процессов на (Ω, \mathcal{F}) со зна-

чениями в фазовом пространстве (X, \mathfrak{M}) , характеризуемый семейством управляемых условных вероятностей $\mu_{t+1}(\cdot | x^t, y^t)$.

Согласно предположению о системе $\{\mu_{t+1}\}$, в классе случайных процессов, образующих управляемый случайный процесс, содержится бесконечное множество элементов. Для того чтобы выделить из этого класса какой-нибудь один случайный процесс, следует назначить способ выбора в каждый момент времени управления (действия) y_t . В простейшем случае этот способ заключается в указании фиксированной последовательности значений управлений $y_0, y_1, \dots, y_t, \dots$, т. е. в задании программы управления — функции времени $y_t = f(t)$. Даже в случае двухэлементного $Y = \{y_1, y_2\}$ существует континuum различных программ.

Правило выбора действия в момент t представляет собой условное распределение на Y

$$F_t(N | x^t, y^{t-1}), \quad N \in \mathfrak{N}, \quad t \geq 1,$$

которое зависит от предшествующих значений процесса и управлений. При вырожденном распределении это рандомизированное правило переходит в детерминированное, изображаемое функцией $f_t(x^t, y^{t-1})$ на $X^t \times Y^t$ со значениями в Y . Всюду в дальнейшем мы без дополнительных оговорок подразумеваем, что распределения $F_t(N | x^t, y^{t-1}), t \geq 1$, при любом $N \in \mathfrak{N}$ и всех t измеримы по совокупности величин (x^t, y^{t-1}) , фигурирующих в условии. Естественно, что в детерминированном случае функции $f(x^t, y^{t-1})$ измеримы на $X^t \times Y^t$.

Обозначим через σ совокупность правила выбора действий. В общем случае $\sigma = \{F_t, t \geq 1\}$, а в детерминированном варианте $\sigma = \{f_t, t \geq 1\}$.

Стратегией управляемого случайного процесса называется совокупность σ правил выбора действий. Каждая стратегия выделяет конкретный случайный процесс из класса, порожденного семейством управляемых условных распределений.

В дальнейшем мы будем рассматривать классы стратегий, обозначаемые $\Sigma = (\sigma)$.

Перечислим некоторые типы стратегий.

Стационарные стратегии образованы из одинаковых правил. Если время t пробегает все целые значения ($\dots, -1, 0, 1, \dots$), то $F_t = F$; если же существует начальный момент ($t=0, 1, 2, \dots$), то, начиная с некоторого t_0 , все правила F_t совпадают и имеют вид $F_t(\cdot | x_{t-h}^{t-1}, y_{t-h}^{t-1})$. Число h назовем глубиной памяти стратегии. Стационарные стратегии, которые являются программными, заключаются в повторении одного единственного действия $y_t = y^* \in Y$.

Марковские стратегии порождены правилами вида $F_t(N | x_t)$, т. е. представляют собой распределения на Y при условии, относящемся лишь к значению процесса в текущий момент времени. В детерминированном случае марковская стратегия есть $\sigma = \{f_t(x)\}$. Марковская стационарная стратегия порождена одной функцией $f(x)$, т. е. во все моменты времени действие избирается по неизменному правилу. Оказывается, что большинство задач теории управляемых процессов решается посредством марковских стратегий, которые оказываются к тому же стационарными.

Отметим два способа формирования (синтеза) стратегии. Первый состоит в том, что заранее для каждого момента t указывается правило выбора действия в виде конкретной функции не более чем $2t$ аргументов. Второй способ делает выбор этой функции зависящим от предшествующего течения процесса. Такой подход естествен, когда используемые функции имеют специальный вид, например, марковский $f_t = f(x_t)$. Предыстория процесса и управления определяют конкретную функцию на X .

Совокупность семейств распределений $\{\mu_t, F_t, t \geq 1\}$ определяет случайный процесс (ξ_t, y_t) в пространстве $X \times Y$, а значит, и вероятностную меру на σ -алгебре пространства элементарных событий (Ω, \mathcal{F}) . Легко построить эту меру на «цилиндрических» множествах

$$(M_{t_1}, N_{t_1}) \times (M_{t_2}, N_{t_2}) \times \dots \times (M_{t_k}, N_{t_k}),$$

$$t_1 < \dots < t_k,$$

а затем обычным методом она продолжается на $X^\infty \times Y^\infty$ — множество всех двумерных последовательностей (x_t, y_t) .

Пусть φ — измеримое отображение $X^t \times Y^t$ на числовую прямую. Тогда $\varphi_t = \varphi(x^t, y^{t-1})$ — скалярный управляемый случайный процесс, вероятностные характеристики которого в принципе вычисляются по совокупности мер $\{\mu_t\}$, $\{F_t\}$ и по функции φ .

Рассмотрим некоторые классы управляемых случайных процессов.

Простейшим классом управляемых случайных процессов являются *процессы с независимыми значениями* (кратко, ПНЗ), у которых семейство управляемых условных вероятностей имеет вид

$$\mu_{t+1}(M | x^t, y^t) \equiv \mu_{t+1}(M | y_t),$$

т. е. в условии фигурирует лишь последнее приложенное управление, но не предыстория самого процесса. В детерминированном варианте такие процессы изображаются однозначной зависимостью состояний ξ_{t+1} от действия y_t :

$$\xi_{t+1} = g_t(y_t).$$

Особую роль играют *однородные процессы с независимыми значениями* (сокращенно ОПНЗ). У них управляемые условные вероятности неизменны во времени, т. е. $\mu_t(\cdot | y) \equiv \mu(\cdot | y)$. Наименование процессов указанного вида связано с тем, что при программной стратегии такой процесс становится последовательностью независимых случайных величин и при этом одинаково распределенных, если $y \equiv \text{const}$. Позднее мы увидим, что именно такой вид имеют стратегии для ОПНЗ в задачах теории управления. Однако усложнение стратегии превращает ОПНЗ в марковский процесс. Действительно, применение марковских стратегий с детерминированным правилом $f_t^*(x_t)$ приводит к условным распределениям $\mu_{t+1}(M | f(x_t))$. Очевидно, стратегии с большей глубиной памяти приводят к сложным марковским процессам.

Функционалы вида $\varphi_t = \varphi(x_{t-l}^t, y_{t-l}^{t-1})$ на траекториях ОПНЗ служат примерами процессов более общей структуры. К последним относятся управляемые марковские процессы, характеризуемые условием

$$\mu_{t+1}(M | x^t, y^t) \equiv \mu_{t+1}(M | x_t, y_t), M \in \mathfrak{M}, t \geq 0.$$

Использование марковских стратегий делает эти процессы марковскими или сложными марковскими, если глубина стратегии фиксирована. Однородность имеет место, если стратегии стационарны, а меры $\mu(M | x, y)$ не меняются со временем.

Частным случаем этих процессов служат управляемые марковские цепи $(X, \mathcal{P}^{(y)}, Y)$, где X не более чем счетное множество состояний, $\mathcal{P}^{(y)} = \| p_{ij}(y) \|$ — стохастическая матрица вероятностей переходов в результате приложения управления $y \in Y = (y_1, \dots, y_k)$, т. е. заданы k таких матриц.

Более сложной природой обладает эволюция «доходов» в управляемой марковской цепи. Этот объект записывается четверкой $(X, \mathcal{P}^{(y)}, R^{(y)}, Y)$, где в дополнение к сказанному ранее фигурируют числовые матрицы $R^{(y)} = \| r_{ij}(y) \|$, элементы коих суть величины доходов, получаемых в результате перехода из x_i в x_j под воздействием управления y . Во всех интересных случаях стратегия для таких цепей оказалась стационарной марковской (и даже детерминированной, т. е. вида $y_t = f(\xi_t)$). Марковская цепь с доходами (и просто — марковская цепь) называется *эргодической*, если при каждой стационарной марковской стратегии получающаяся цепь эргодична в том смысле, который был определен ранее, в § 2.

Другие примеры управляемых марковских процессов зададим уравнениями. В случае вырожденных процессов это обычно разностные уравнения

$$\xi_{t+1} = g(\xi_{t-1}^t, y_{t-1}^t),$$

а в стохастическом варианте

$$\xi_{t+1} = g(\xi_{t-1}^t, y_{t-1}^t, \zeta_t),$$

где ζ_t — последовательность независимых случайных величин.

§ 4. Цели управления

Управление эволюцией объекта должно привести к тому, чтобы объект обладал некоторыми желаемыми свойствами. Эти предписываемые объекту свойства являются целью управления. Сделать понятие «цели» одновременно

и формальным, и универсальным (охватывающим все или почти все интересные случаи) затруднительно. Поэтому ниже описаны примеры характерных целей, тех, к достижению которых мы будем стремиться в дальнейшем.

Формулировки рассматриваемых нами целей относятся к свойствам функционалов на траекториях. Применимтельно к стохастическим объектам, т. е. с невырожденными распределениями, приходится рассматривать математические ожидания функционалов. Для вычисления их следует предварительно избрать стратегию σ , с помощью которой на траекториях случайного процесса задается вероятностная мера. Выпишем сначала математическое ожидание функционала $\varphi_t = \varphi(x^t, y^{t-1})$ в случае программной стратегии $\sigma = \{f\}$; здесь $f(t) = y_t$ означает действие (управление), предписываемое в момент t этой стратегией (здесь и ниже мы допускаем, что указываемые интегралы существуют)

$$\begin{aligned} E_{\varphi_{t+1}} = W(y^t) &= \int_{x^t} \varphi(x_1, \dots, x_{t+1}; y_0, y_1, \dots, y_t), \\ \mu(dx_{t+1} | x^t, y^t) \mu(dx_t | x^{t-1}, y^{t-1}) \dots & \\ \dots \mu(dx_2 | x_1, y_0, y_1) \mu(dx_1 | y_0). \end{aligned}$$

Таким образом, $E_{\varphi_{t+1}}$ оказывается функцией предшествующих управлений.

Теперь рассмотрим детерминированную стратегию $\sigma = \{f_t\}$, тогда математические ожидания E_{φ_t} образуют не функциональную последовательность, а числовую. Для построения ее исключим y_1, \dots, y_t из числа аргументов функционала φ и из функций $f_n(x^n, y^{n-1})$. С этой целью следует воспользоваться цепочкой равенств

$$\begin{aligned} y_t &= f_t(x^t; y^{t-1}), \quad y_{t-1} = \\ &= f_{t-1}(x^{t-1}, y^{t-2}), \dots, y_3 = f_3(x_1, x_2, x_3; y_0, y_1, y_2), \\ y_2 &= f_2(x_1, x_2; y_0, y_1), \quad y_1 = f_1(x_1, y_0), \end{aligned}$$

которые сохраняют только начальное управление y_0 . В результате приходим к системе функций, определяющих управления $y_t = f_t(x_1, \dots, x_t; y_0)$, $t \geq 1$. С их по-

мощью математическое ожидание $E_{\sigma} \varphi_t$ при детерминированной непрограммной стратегии σ выражается формулой

$$\begin{aligned} E_{\sigma} \varphi_t = W_{\sigma, y_0}(t) &= \\ &= \int_{X^t} \varphi(x_1, \dots, x_t; y_0, \tilde{x}_1, \dots, \tilde{x}_{t-1}) \times \\ &\quad \times \mu(dx_t | x^{t-1}, \tilde{x}^{t-1}) \dots \mu(dx_2 | x_1; y_0, \tilde{x}_1) \mu(dx_1 | y_0). \end{aligned}$$

Остается, наконец, общий случай рандомизированной стратегии. Порождаемая ею мера на множестве траекторий случайного процесса приводит к следующему выражению:

$$\begin{aligned} E_{\sigma} \varphi_t = W_{\sigma}(t) &= \int_{X^t \times Y^t} \varphi(x_1, \dots, x_t; y_0, y_1, \dots, y_{t-1}) \times \\ &\quad \times \prod_{j=1}^t \mu(dx_j | x^{j-1}, y^{j-1}) F(dy_{j-1} | x^{j-1}, y^{j-2}). \end{aligned}$$

Множество Σ допустимых стратегий подчинено требованию, чтобы при всех t существовали (и были конечными) математические ожидания $E_{\sigma}(t)$.

Удобна следующая терминология: значение функционала φ_t называется *выигрышем* в момент времени t , а его математическое ожидание — *средним выигрышем*.

В случае программной стратегии средний выигрыш $W(y^t)$ представляет собой функцию на Y^{t+1} . Из принятого в § 3 допущения, что управляемые условные вероятности измеримы относительно фигурирующих в условии совокупности прошлых действий, вытекает

Теорема 1. *Функции $W(y^t)$ при каждом t измеримы в области своего определения.*

Если пространство управлений Y топологическое, то справедлива

Теорема 2. *Если условные вероятности $\mu(M | x^t, y^t)$ непрерывны по y^t при любых x^t и $M \in \mathfrak{M}$ и интегралы, определяющие $W(y^t)$, сходятся равномерно, то средние выигрыши $W(y^t)$ непрерывны по совокупности аргументов.*

Доказательства обеих теорем проводятся стандартным образом и потому опускаются.

Среди различных типов функционалов особое значение имеют функционалы вида $\varphi_t = \varphi(\xi_t)$. Их роль объясняется тем, что во многих случаях о текущем состоянии ξ_t процесса судят по результату измерения какой-то численной характеристики состояния. Математическое ожидание этого функционала в случае программной стратегии записывается так:

$$\mathbf{E}\varphi_t = W(y^{t-1}) = \int_X \varphi(x) \mu^{(t)}(dx | y^{t-1}),$$

где для любого $M \in \mathfrak{M}$ мера $\mu^{(t)}$ имеет вид

$$\begin{aligned} \mu^{(t)}(M | y^{t-1}) &= \int_{X^{t-1}} \mu(M | x^{t-1}; y^{t-1}) \mu(dx_{t-1} | x^{t-2}; y^{t-2}) \dots \\ &\quad \dots \mu(dx_2 | x_1; y_0, y_1) \mu(dx_1 | y_0). \end{aligned}$$

Не составляет труда выписать соответствующее выражение и при иных стратегиях, включая общий случай рандомизированных правил. Отметим, что применительно к простейшему классу процессов — ОПНЗ эта формула существенно упрощается и оказывается следующей:

$$\mathbf{E}\varphi_t = W(y_{t-1}) = \int_X \varphi(x) \mu(dx | y_{t-1}).$$

Средний выигрыш в каждый момент времени для таких процессов зависит лишь от приложенного на предыдущем такте управления. Это обстоятельство делает проблемы управления процессами типа ОПНЗ относительно простыми.

Обратимся к перечислению характерных рассматриваемых целей управления. Их объединяет следующее: они относятся к свойствам процессов, которые должны выполняться во все моменты времени, начиная с некоторого (начального либо какого-нибудь последующего). Этим исключаются «локальные» цели вроде такой: траектория процесса хотя бы раз должна пройти через данную точку $x' \in X$ (или оказаться в ней в момент t_0). Иными словами, цель управления, как и процесс, должна быть «необрывающейся», т. е. требующей неограниченного продолжения управления процессом.

Задача о «выполнении плана» состоит в требовании, чтобы средние выигрыши удовлетворяли неравенствам

$$a(t) < W(t) < b(t), \quad t \geq t_0,$$

где $a(t)$ и $b(t)$ — заданные числовые последовательности. Более сильная форма этой задачи требует выполнения равенств $W(t)=c(t)$. Если допустимо ограничиваться программными стратегиями, эта цель решается построением последовательности действий $y_0, y_1, \dots, y_t, \dots$ такой, что начиная с момента t_0 , $a(t) < W(y^{t-1}) < b(t)$.

В применении к ОПНЗ с функционалом $\varphi(\xi_t)$ задача о выполнении плана решается программной стратегией $\sigma = \{f(t), t \geq 1\}$, обеспечивающей выполнение неравенств

$$a(t) < W(y_t) < b(t).$$

Задача о выполнении равенства $W(y_t) = W_0$ решается стационарной программной стратегией $y_t = y^0$, где управление y_0 — корень уравнения $W(y) = W_0$.

Пусть теперь на траекториях процесса заданы $m \geq 1$ функционалов $\varphi^{(1)}, \dots, \varphi^{(m)}$. Требуется, чтобы при всех $t \geq t_0$ выполнялись включения

$$(W^{(1)}(t), \dots, W^{(m)}(t)) \subset G,$$

где $W^{(j)}(t) = E^{(j)} \varphi_t$, $j = \overline{1, m}$, а G — область m -мерного евклидова пространства. К такой или подобной форме сводится большинство известных целей управления. В качестве примера отметим задачу об «устойчивости движения» случайного процесса в гильбертовом пространстве, которая требует, чтобы начиная с некоторого момента

$$E \|\xi_t\| < r, \quad r > 0.$$

Если распределения управляемого процесса вырождены (т. е. его можно трактовать как детерминированный), любую цель можно представить в форме неравенства. Для этого зададим такой функционал:

$$\varphi_t = \varphi(\xi_t) = \begin{cases} 1, & \xi_t \in U_t, \\ -1, & \xi_t \notin U_t, \end{cases}$$

где U_t — множество траекторий длины t , обладающих желаемым свойством. Остается потребовать, чтобы при всех $t \geq t_0$ было $\varphi_t > 0$. Аналогичное представление целей для произвольных (т. е. случайных невырожденных) процессов возможно не всегда. Цели типа равенств означают в детерминированном случае, что процесс должен с некоторого момента удовлетворять заданному разностному уравнению $h(\xi_t, \xi_{t+1}, \dots, \xi_{t+n})=0$.

Задачи о максимизации выигрыша, по-видимому, наиболее распространены и имеют много разных форм.

Задан класс K управляемых случайных процессов с одинаковыми пространствами: фазовым X и пространством управлений Y . Пусть Σ — множество допустимых стратегий для всех процессов из K . Обозначим начальный отрезок длины t (т. е. совокупность первых t правил) стратегии σ через σ_t и положим $\bar{W}(t) = \sup_{\sigma_t \in \Sigma} W_\sigma(t)$. Введем величину

$$\bar{W} = \lim_{\overline{t \rightarrow \infty}} \bar{W}(t).$$

Таков, по определению, максимальный предельный средний доход.

Цель управления — максимизация выигрыша в *сильном смысле* представляет собой задачу синтеза стратегии, которая для любого процесса из K обеспечивает неравенства ($\varepsilon > 0$)

$$\frac{1}{t} \sum_1^t \varphi_j > \bar{W}(t) - \varepsilon$$

либо

$$\frac{1}{t} \sum_1^t \varphi_j > \bar{W} - \varepsilon$$

при всех $t > \tau_\varepsilon(\omega)$, где $P(\tau_\varepsilon(\omega) < \infty) = 1$. В этих неравенствах ε либо фиксировано, тогда говорят о цели — ε -оптимальности, либо оно произвольно, тогда целью

служит *асимптотическая оптимальность*. Последняя эквивалентным образом записывается в виде пределов

$$\lim_{t \rightarrow \infty} \left[\frac{1}{t} \sum_1^t \varphi_j - \bar{W}(t) \right] = 0, \quad \lim_{t \rightarrow \infty} \frac{1}{t} \sum_1^t \varphi_j = \bar{W},$$

которые можно понимать в разных смыслах: по вероятности, в среднем квадратическом, с вероятностью 1 и т. д.

Цели в слабом смысле формулируются с помощью средних выигрышей $W_\sigma(t)$ при заданном функционале φ_t . Отыскиваются стратегии, обеспечивающие при всех $t \geq t_0$ для любого процесса из класса K

$$W_\sigma(t) > \bar{W}(t) - \varepsilon, \quad \varepsilon > 0,$$

либо

$$W_\sigma(t) > \bar{W} - \varepsilon.$$

Как и ранее, здесь число ε либо фиксировано (ε -оптимальность в слабом смысле), либо произвольно (асимптотическая оптимальность в слабом смысле).

Для некоторых классов управляемых случайных процессов сильные и слабые цели равносильны. Так обстоит дело, например, для ОПНЗ и марковских цепей с доходами.

Иногда приходится ограничиваться еще более слабыми целями. Например, требуется, чтобы для каждого процесса из K при всех $t \geq t_0$ выполнялись неравенства

$$\frac{1}{t} \sum_{s=1}^t W_\sigma(s) > \bar{W} - \varepsilon, \quad \varepsilon > 0,$$

или их модификация с $\bar{W}(t)$. Существуют такие процессы, по отношению к которым недостижимы более сильные цели, чем приведенная.

Обратимся к задачам типа условного экстремума, в которых на траекториях процесса определены $m \geq 2$ функционалов $\varphi^{(1)}, \dots, \varphi^{(m)}$ с конечными математическими ожиданиями $W^{(1)}, \dots, W^{(m)}$. В слабой постановке цель управления заключается в максимизации $W^{(1)}$ при соблюдении ограничений $W^{(j)} > 0, j=2, \dots, m$. Максимиза-

ция $W^{(1)}$ означает, как и выше, построение такой стратегии σ , что $W_{\sigma}^{(1)}(t) > \bar{W}^{(1)} - \varepsilon$ с фиксированным либо произвольным ε . Один из вариантов постановки сильной цели заключается в следующем:

$$\frac{1}{t} \sum_1^t \varphi^{(1)}(\xi^t) > \bar{W}^{(1)} - \varepsilon, \quad t > \tau_{\varepsilon},$$

при условиях на $W_{\sigma}^{(j)}, j \geq 2$. Нетрудно представить себе другие варианты постановок сильных целей.

Полезно отметить связь между целями оптимизационными и целями в виде неравенств. Мы видим, что первые из них имеют вид неравенств, относящихся либо к самим функционалам, либо к их математическим ожиданиям. Справедливо обратное, цели, сформулированные как обеспечение выполнения неравенств, можно переформулировать в оптимизационные. Это совсем просто в детерминированном случае: достаточно вместо предложенного функционала ψ , который следует сделать положительным, построить новый φ . Его мы выберем равным 1 на множестве тех траекторий, на которых $\psi > 0$, и равным 0 на дополнительном множестве. Теперь цель управления — максимизировать функционал φ .

§ 5. Постановка задач в классической теории управления

Предмет классической теории управления формулируется на развитом в этой главе языке как синтез стратегии для управляемого случайного процесса. Всегда считают точно известным математическое описание объекта управления. В детерминированном случае задана не только структура «уравнений движения», но и все фигурирующие в них функции и параметры, а в стохастическом — система управляемых условных вероятностей (или какой-нибудь иной способ вероятностного задания процесса), известных полностью и точно. По этим данным вычисляется* приходящая к цели стратегия. Находится она до начала фактического воздействия на объект и в процессе функционирования изменения в нее не вносятся (не считая, ра-

зумеется, устранения замеченных ошибок вычислений). Порядок применения слагающих стратегию правил (распределений F_t , в рандомизированном варианте и функций f_t , в детерминированном) назначают заранее.

Поясним сказанное несколькими примерами. Сначала пусть объектом управления служит скалярный ОПНЗ ξ_t . Допустим, что цель управления относится к среднему выигрышу в единицу времени $E \xi_t = W(y)$. Если требуется выдерживать средний выигрыш на заданном уровне W_0 , следует решить уравнение

$$W(y) = W_0$$

и затем применить стационарную программную стратегию $y_t \equiv y_0$ — корень этого уравнения. Если цель состоит в максимизации среднего выигрыша, то каким-либо методом находим точку глобального максимума

$$W(y^0) = \max_{y \in Y} W(y)$$

и снова пользуемся стратегией $y_t \equiv y^0$.

Пусть управлению подлежит конечная эргодическая марковская цепь с доходами $(X, \mathcal{P}^{(y)}, R^{(y)}, Y)$ и следует обеспечить максимум предельного дохода за один такт. Можно показать, что для достижения этой цели достаточно ограничиться марковскими стратегиями $y=f(x)$. Такую стратегию удобно изобразить $|X|$ -мерным вектором $y=(y_1, \dots, y_{|X|})$, где $y_j=y(x_j)$ — управление, которое сопоставлено состоянию X_j . Различных стратегий $|Y|^{|X|}$ и их можно перенумеровать y_1, y_2, \dots

По условию рассматриваемая цепь эргодическая и, следовательно, при каждой стратегии y существует (и не зависит от начального состояния) предельное распределение $\pi(y)=(\pi_1(y), \dots, \pi_{|X|}(y))$ на множестве состояний X . Интересующий нас предельный средний доход в единицу времени выражается следующим образом:

$$W(\bar{y}) = \sum_{j=1}^{|X|} \pi_j \sum_{i=1}^{|X|} p_{ij}(y(x_i)) r_{ij}(y(x_i)),$$

где $y(x_i)$ означает действие, сопоставленное стратегией y состоянию цепи x_i . При малых $|X|$ и $|Y|$ не составляет труда перебрать $W(y)$ для всех стратегий и найти «оптимальную» y_{opt} , максимизирующую функцию $W(y)$. Существуют также простые итеративные процедуры вычисления оптимальной стратегии. Применение таких методов приводит к искомой стратегии, которая используется с начала управления цепью.

Если в обоих рассмотренных примерах управляемый объект описан неточно (это относится либо к мере $\mu(M|y)$, либо к совокупности матриц $\mathcal{P}^{(y)}, R^{(y)}$), то вычисленная стратегия может оказаться не только не оптимальной, но даже отдаляющей от поставленной цели. Однако уточнение описания объекта не предусматривается классической теорией управления. Поэтому сколь угодно длительное функционирование «оптимальной» системы, синтезированной на основе принципов этой теории, не приведет к желаемой цели.

ГЛАВА II

АДАПТИВНЫЕ СИСТЕМЫ УПРАВЛЕНИЯ

§ 1. Определение адаптивных систем управления

Рассматриваются класс K управляемых случайных процессов и класс Φ функционалов на траекториях процессов из K . Задано множество $\Sigma = \{\sigma\}$ допустимых стратегий для всех процессов из K , порождающее вероятностные меры на пространстве элементарных событий. Сформулирована цель управления, относящаяся к произвольной паре (ξ, φ) из множества $(K \times \Phi)$ и достижимая на всем этом множестве. Предполагается, что в классе K содержится по меньшей мере два процесса (различные системы управляемых условных вероятностей), а все функционалы из Φ имеют конечные математические ожидания для каждого процесса из K , т. е. по мерам, индуцированным стратегиями из множества Σ . Цель управления сформулирована в терминах этих математических ожиданий.

Адаптивной системой управления называется стратегия, которая приводит к цели управления для всякой пары $(\xi, \varphi) \in (K \times \Phi)$ за конечное (с вероятностью единица) время.

Определение подразумевает, что свойство адаптивности управляющей системы относится к классу (объектов и функционалов). Ниже всегда будет ясно без дополнительных оговорок, какие классы имеются в виду. Существенно уяснить, что адаптивная система «не знает» процесс, которым она управляет. В процессе функционирования она может вести оценку характеристик управляемого ею процесса, находить с возрастающей точностью его «уравнения движения», однако не обязательно управление совмещено с оцениванием свойств объекта (или, как иногда говорят, решением задачи идентификации).

Адаптивная система управления не является стационарной стратегией. Правила выбора действий подбира-

ются в ходе управления (а не назначаются заранее), их перебор осуществляется на основании поступающих от объекта сигналов. Таким образом, значения процесса, текущие и прошлые, не только порождают действия системы управления, но и смену правил их выбора. Подробнее эти вопросы мы обсудим в следующем параграфе.

В реальных ситуациях управляемый случайный процесс ξ_t нередко недоступен прямой фиксации и измерениям. Более того, существуют такие явления, в которых не вполне ясно, каково фазовое пространство, а это на практике приводит к тому, что неизвестно, какие измерительные приборы необходимы. Поэтому о течении управляемого процесса ξ_t приходится судить по значениям наблюдаемого процесса η_t , связанного с ξ_t так, что восстановить по нему ξ_t нельзя. Приведем два примера наблюдаемых процессов, встречающихся в задачах управления векторными марковскими процессами вида

$$\xi_{t+1} = A(y_t)\xi_t + B(y_t)\zeta_t,$$

где ζ_t — последовательность независимых случайных векторов, $A(y_t)$ и $B(y_t)$ — матрицы. В одном из часто изучаемых вариантов наблюдают числовую последовательность

$$\eta_t = (\mathbf{c}, \xi_t)$$

— скалярное произведение постоянного вектора \mathbf{c} на вектор текущих значений процесса, т. е. линейную комбинацию компонент вектора ξ_t . В другом варианте наблюдается процесс (\mathcal{D} - и \mathcal{E} -матрицы)

$$\eta_{t+1} = \mathcal{D}(\xi_t)\eta_t + \mathcal{E}\zeta_t,$$

стохастически связанный с управляемым.

Наконец, наблюдаемым процессом может быть функционал $\varphi_t = \varphi(\xi^t, y^{t-1})$, к математическому ожиданию которого относится цель управления. В этом и других случаях правила выбора действий зависят от значений не управляемого, а наблюдаемого процесса. Так распределения вероятностей F_t принимают вид $F_t(N | \eta^t, y^{t-1})$.

В дальнейшем мы всегда будем полагать, что управляемый процесс совпадает с наблюдаемым. Ясно, что это ни в малой степени не ограничивает общности. Кроме того, множество Φ будет содержать один элемент.

§ 2. Обучаемые системы

Введем понятие обучаемой системы, которая взаимодействует с любым управляемым случайным процессом, имеющим фазовое пространство X и пространство управлений Y .

Зададим правило управления F , т. е. вероятностное отображение $X^l \times Y^{l-1}$ на Y , l — целое число. В вырожденном случае, когда правило детерминировано, действие y_t есть значение функции $f(x_{t-l}^t, y_{t-l}^{t-1}) = y_t$. Стационарную стратегию, порожденную распределением F или функцией f , реализует элементарная управляющая система U_F (или U_f), т. е. объект

$$U_F = (X, F, Y).$$

Входными сигналами U_F служат значения управляемого (и наблюдаемого) процесса ξ_t , а выходными — управление $y \in Y$. Вычисление действий в смежные моменты времени t и $t+1$ основано соответственно на предыстории $(x_{t-l+1}, \dots, x_t; y_{t-l+1}, \dots, y_{t-1})$ и $(x_{t-l+2}, \dots, x_t, x_{t+1}; y_{t-l+2}, \dots, y_{t-1}, y_t)$. О таких системах говорят, что они имеют память глубины l .

По определению, элементарная управляющая система U_F управляет процессом ξ_t , если после ее действия y_{t-1} процесс принимает значение x_t из множества $M \in \mathfrak{M}$ с вероятностью $\mu_t(M | x^{t-1}, y^{t-1})$. В тот же момент ξ_t попадает на вход U_F и вызывает, согласно правилу F , очередное действие y_t управляющей системы.

Обозначим через D_l совокупность всех условных распределений на Y вида

$$F(N | x^l, y_0^{l-1}),$$

измеримых на $X^l \times Y^{l-1}$, и положим $D_\infty = \bigcup_l D_l$. Выберем в D_∞ непустое подмножество $D \subseteq D_\infty$. Оно служит множеством допустимых правил, фигурирующих в рассматриваемых стратегиях.

Через \tilde{U} обозначим множество всех элементарных управляющих систем, которое сопоставлено множеству D

допустимых правил

$$(U_F \in \tilde{U}) \Leftrightarrow \{U_F = (X, F, Y), F \in D\}.$$

Это множество символически записывается в виде $\tilde{U} = (X, D, Y)$.

Пусть на управляемом случайному процессе ξ_t заданы $m \geqslant 1$ функционалов, точнее, измеримое отображение $\zeta_t: X^t \times Y^t \rightarrow R_m$, m -мерное евклидово пространство. Назовем это отображение *статистикой процесса* ξ_t . В простейших случаях это скалярная величина $\zeta_t = \Psi(\xi_t)$ или даже, если X — множество на числовой прямой, $\zeta_t = \xi_t$.

Символом T воспользуемся для обозначения отображения множества D в себя, т. е. $T: D \rightarrow D$. Рассматривается двухпараметрическое семейство $T_{\zeta_t, t}$ таких отображений, параметрами служат статистика ξ и время t . Операторы $T_{\zeta_t, t}$ естественно также считать определенными на множестве \tilde{U} управляющих систем. Зависимость этого семейства от ζ означает, что вид оператора T в каждый момент времени определяется предысторией процесса (и совершенными в прошлом действиями).

Обучаемой системой L называется объект

$$L = [\tilde{U}, T_{\zeta_t, t}].$$

Выбранный термин — «обучаемая система» — обусловлен тем, что объект L является обобщением «стохастических моделей обучаемости» (о них будет сказано ниже), наименование которых стало общепринятым.

Функционирование обучаемой системы протекает следующим образом: получив на входе сигнал x_t , она вырабатывает очередное правило $F_t = T_{\zeta_t, t} F_{t-1}$, которое определяет выходной сигнал (действие). Следующий входной сигнал x_{t+1} приводит к аналогичному циклу. Сказанное подразумевает, что обучаемая система обладает «памятью», в которой хранится необходимое число прошлых действий (y_t) и откликов (ξ_t) на них. В общем случае памятью служит бесконечное произведение пространств $X^\infty \times Y^\infty$, но достаточно рассматривать $X' \times Y'^{l-1}$ (или даже его подмножество), если объем памяти системы ограничен сверху числом l . В частности, так обстоит дело в задачах управления марковскими процессами (и, разумеется, ОПНЗ).

Смысл понятия «обучаемая система управляет процессом» очевиден. Совокупность из обучаемой системы L и управляемого ею процесса ξ обозначим символом $L \otimes \xi$.

Итак, обучаемая система реализует стратегию управления случайным процессом. В каждый момент времени она совершает действия в зависимости от предыстории: уже выбранных действий и откликов на них управляемого процесса. Реализуемая такой системой стратегия, вообще говоря, нестационарная.

Рассмотрим взаимодействие системы L и процесса ξ_t . При появлении на входе L последовательности сигналов ξ_1, \dots, ξ_m начальное правило выбора действий f_0 переходит в $F_m = T_{\xi_m, m} \dots T_{\xi_1, 1} F_0$ — m -кратную итерацию операторов $T_{\xi_i, i}$. Случайность входных сигналов влечет за собой случайность операторов. Следовательно, последовательность правил F_t образует случайный процесс. Иными словами, управляемый случайный процесс ξ_t порождает на множестве D случайное блуждание.

Рассмотрим пример обучаемой системы. Процесс ξ_t , с которым она будет взаимодействовать, имеет конечное фазовое пространство $X = \{x_1, \dots, x_m\}$, элементы которого называются стимулами, и конечное пространство управлений $Y = \{y_1, \dots, y_n\}$, называемых реакциями. Появление стимулов регулируется условными вероятностями $\mu(x | y)$, т. е. ξ_t представляет собою ОПНЗ.

Наименование «стохастическая модель обучаемости» — МБМ (модель Буша — Мостеллера) утвердилось за следующей обучаемой системой, которая явилась первой (пока единственной) попыткой количественного описания явлений приспособления животных к внешним условиям. Ее множествами входных и выходных сигналов являются соответственно X и Y . Множеством допустимых правил D служит вероятностный $(k - 1)$ -мерный симплекс $D = \left\{ (p_1, \dots, p_k) : \sum_1^k p_j = 1, p_j \geq 0 \right\}$ с метрикой, индуцированной содержащим симплекс евклидовым пространством. Его элементы — стохастические векторы $p = (p_1, \dots, p_k)$ — называют *поведением* МБМ. Компонента p_j этого вектора означает вероятность появления реакции y_j . Поступающий

в ответ на реакцию стимул x от процесса ξ (его в теории обучения называют «средой», «внешним миром», «экспериментатором») вызывает трансформацию текущего значения поведения p . Принята следующая форма преобразования T_x :

$$T_x p = \alpha_x p + (1 - \alpha_x) q_x,$$

где $\alpha_x \in [0, 1]$, а вектор $q_x = (q_1(x), \dots, q_k(x))$ стохастический. Легко проверить, что $T_x^n p = \alpha_x^n p + (1 - \alpha_x^n) q_x$. Отсюда следует, что вектор q_x является неподвижной точкой преобразования T_x , а число α_x определяет скорость приближения $T_x^n p$ к q_x . Поэтому, если от среды на вход МБМ достаточно долго поступает лишь стимул x , поведение МБМ станет практически неотличимым от q_x .

При появлении на входе МБМ стимулов от среды поведения p_t меняются со временем и образуют случайный процесс. Легко видеть, что при сделанном предположении о среде p_t является марковским процессом. Известно, что распределения вероятностей поведений сходятся слабо к предельному.

Обратим внимание на то, что в изложенной концепции взаимодействия среды и МБМ отсутствует понятие цели: МБМ получает из среды стимулы и откликается на них реакциями. «Хороши» последние или «плохи» — этот вопрос не ставится в теории стохастических моделей обучаемости.

Понятие обучаемой системы тесно связано с понятием автомата. Перед тем как это установить, введем необходимые обозначения. Мы рассматриваем вероятностные автоматы Мура, т. е. объекты

$$A = (X, S, Y; s_0; \Pi(x), q).$$

Здесь X и Y — конечные множества входных и выходных сигналов, S — не более чем счетное множество состояний, $s_0 \in S$ — начальное состояние *). Символ $\Pi(x)$ означает семейство стохастических матриц $\|\pi_{ij}(x)\|$ вероятностей переходов $\pi_{ij}(x) = P(s_{t+1}=s_j | s_t=s_i, x_{t+1}=x)$, а $q(y|s) =$

*.) Иногда задают распределение вероятностей начального состояния, но мы ограничились вариантом, когда оно вырождено. Заметим, что встречаются неинициальные автоматы, у которых не выбрано начальное состояние.

$=\mathbf{P}(y_t=y|s_t=s)$ (условное распределение на Y) функцию выходов. Если все перечисленные распределения вырождены, автомат называют *детерминированным*. В случае вырожденности функции выходов q множество состояний S разлагается в сумму непересекающихся множеств $S^{(j)}$, причем всем состояниям из $S^{(j)}$ отвечает один и тот же выходной сигнал y_j .

Иногда бывает необходимо рассматривать автоматы с переменной структурой. Так именуют автоматы, у которых функции переходов и выходов явным образом зависят от времени $\Pi=\Pi(x, t)$ и $q=q(t)$. Нелишне заметить, что усложнением состояния автомата — введением в него времени, т. е. рассмотрением вместо множества $S=\{s\}$ произведения $S \times T=\{(s, t)\}$ (где $T=(0, 1, 2, \dots)$ — множество значений времени), — автоматы с переменной структурой сводятся к обычным автоматам, но, как правило, с бесконечным множеством состояний.

Покажем, что обучаемая система представима обобщенным автоматом, т. е. автоматом с бесконечным числом состояний.

Запишем обучаемую систему $L=[\tilde{U}, T_{\zeta, t}]$ в более подробной форме

$$L=(X, S, Y; T_{\zeta, t}).$$

Здесь X и Y — множества входных и выходных сигналов, $S=(F, R_m, X^\infty \times Y^\infty)$ — множество состояний (входящие в тройку символы уже были ранее определены). Функция переходов автомата должна указывать, как преобразуются правила выбора действия, статистика ζ и содержимое памяти. Преобразование правил задается семейством $T_{\zeta, t}$, способы трансформации статистики и памяти опущены в приведенной записи L , так как подразумевается, что эта переработка осуществляется в соответствии с видом статистики и характером памяти в каждом такте функционирования. Если все три преобразования неизменны во времени (в частности, $T_{\zeta, t} \equiv T$), перед нами ситуация, относящаяся к обычному пониманию автоматов. Если же хотя бы одно меняется со временем, то следует говорить об автоматах с переменной структурой.

Функция выходов обучаемой системы L в каждый момент времени определена текущим состоянием, правилом f_t и содержимым памяти (x^t, y^{t-1}) . Это свойство системы L идентично муровскому свойству автоматов.

Из сказанного вытекает, что обучаемую систему действительно допустимо трактовать как автомат, понимаемый в обобщенном смысле при бесконечных X, Y, S .

В качестве примера представим МБМ в виде автомата. У МБМ входными сигналами служат стимулы из множества X , а выходными — реакции из Y . Множеством состояний S является $(k-1)$ -мерный симплекс, мощность которого — континуум. Функция переходов детерминированная и представляет собой совокупность из преобразований T_{x_1}, \dots, T_{x_k} , индексы которых — стимулы $x \in X$. Статистикой, которая определяет это семейство операторов, служит номер поступающего из среды стимула. Глубина памяти равна единице. Функция выходов является распределением вероятностей на множестве реакций Y , оно совпадает с текущим состоянием МБМ. Начальное состояние не фиксировано. Таким образом, МБМ оказывается обобщенным неинициальным автоматом Мура с вероятностной функцией выходов.

Сложность этого автомата заставляет подчеркнуть достоинства тех обучаемых систем, которые изображаются автоматами в обычном смысле:

- 1) относительная простота исследования подобных систем;
- 2) возможность реализации (физической и технической: либо в виде программ для ЦВМ, либо в виде схемы, используя набор соответствующих функциональных элементов, линии задержки и, в вероятностном варианте, генераторы случайных величин).

Перейдем теперь к конструктивному определению понятия «адаптивная система».

Снова перед нами классы K — управляемых случайных процессов и Φ — функционалов на траекториях процессов из K . Как и ранее, считаем, что каждый функционал имеет конечное математическое ожидание для любого процесса и при любой стратегии из множества Σ допустимых стратегий. Пусть сформулирована цель управления, достижимая для всякой пары $(\xi, \varphi) \in K \times \Phi$.

Адаптивной системой называется обучаемая система, которая для любых $(\xi, \varphi) \in K \times \Phi$ приводит к достижению цели за конечное (с вероятностью 1) время.

Вспоминая сказанное ранее о случайному блуждании по множеству допустимых правил F при взаимодействии обучаемой системы L с процессом ξ_t , мы теперь можем отметить «направленный» характер этого блуждания, когда L является адаптивной системой. Рассмотрим примеры блужданий, приводящих к ε -оптимальности и асимптотической оптимальности (безразлично, в слабом или сильном смысле). В множестве правил D выделим подмножество $D(\varepsilon)$ тех правил, неограниченное повторение которых обеспечивает получение выигрыша, отличающегося от максимального менее чем на ε для управляемого системой процесса. Один из способов достижения ε -оптимальности заключается в том, чтобы обучаемая система в течение подавляющей доли времени пользовалась правилами из $D(\varepsilon)$, а другой требует, чтобы $D(\varepsilon)$ было поглощающим множеством блуждания. Обеспечение ε -оптимальности (в каком-нибудь смысле) для любого $\varepsilon > 0$ означает, что необходимо строить семейство L_ε обучаемых систем, зависящих от ε как от параметра. Будем называть такие семейства ε -оптимальными семействами. Отметим, что «предельная» система L_0 может не обладать никаким полезным свойством.

Асимптотическую оптимальность обеспечивают, среди прочих методов, следующие: а) движение по ε -оптимальным правилам (для данного управляемого процесса) с приближением к оптимальному (впрочем, последнее не обязательно принадлежит множеству D); б) доля времени, в течение которого система пользуется оптимальными правилами, стремится к 1 с ростом времени.

Из сказанного здесь очевидна возможность перефразировки исходной цели (относительно максимизации среднего выигрыша) в цель блуждания по множеству D . Так вместо определенной выше асимптотической оптимальности может оказаться удобным говорить, как о цели, о «стремлении правил выбора действий к оптимальному» и эту последнюю именовать асимптотической оптимальностью. Позднее мы не раз встретимся с подобными пере-

формулировками, позволяющими иногда свести исходную задачу к более удобному для анализа виду.

Синтез адаптивной системы, обеспечивающей достижение цели на классе $K \times \Phi$, означает организацию надлежащего случайного блуждания по множеству D . Среди возможных методов построения таких блужданий особенно прост в принципиальном отношении и поэтому популярен следующий: в ходе управления объектом осуществляется оценка его динамических характеристик, т. е. системы условных распределений или уравнений движения, или входящих в них параметров. На основании оценок вычисляется приводящая к цели стратегия (здесь используются подходы классической теории управления). С течением времени оценки уточняются и приближаются к истинным значениям оцениваемых величин, а потому и вычисляемые стратегии сходятся к «оптимальной». Этот метод связывает теорию адаптивных систем с математической статистикой (либо теорией идентификации).

Функционирование адаптивных систем в терминах математической статистики и теории игр можно формулировать как процесс принятия решений. Отличие его от процедур математической статистики заключается в следующем. Статистические выводы всегда основаны на выборках конечного объема и с положительной вероятностью сопровождаются ошибками. Напротив, «решения» адаптивных систем, принимаемые спустя конечное время τ , безошибочны. Если имеется в виду сильная цель — максимизация выигрыша, время τ оказывается случайной величиной, которая с вероятностью 1 конечна. Из смысла этой величины (при всех $t > \tau$ выполняется неравенство $\frac{1}{t} \sum_{j=1}^t \varphi_j > \bar{W} - \varepsilon$) непосредственно вытекает, что она зависит от будущего поведения управляемого процесса и адаптивной системы, т. е. является немарковским моментом. Отсюда следует, что ни в какой момент деятельности адаптивной системы нельзя быть уверенным, что избранные ею правила из D оптимальны или же близки к ним. В системах, преследующих слабые цели, интересующее нас неравенство наступает, начиная с неслучайного момента времени t_ε (§), зависящего в общем случае от управ-

ляемого процесса и поэтому нам неизвестного. Появление немарковских случайных моментов — характеристика адаптивных систем — является как бы «расплатой» за неопределенность задания объектов управления.

§ 3. Ассоциированные марковские процессы

Объекту $L \otimes \xi$, означающему взаимодействующие обусловленную систему и управляемый ею процесс, мы сопоставим здесь марковский процесс (C, \mathcal{P}) с множеством состояний C и переходной функцией \mathcal{P} . Выясним строение этого процесса.

Вид элементов множества C зависит от глубины последействия управляемого процесса. В общем случае, когда управляемые условные вероятности $\mu_t(\cdot | x^{t-1}, y^{t-1})$ зависят от всей предыстории, состояниями множества C служат наборы

$$c = (x^t; y^{t-1}; F_{t-1}), \quad t = 1, 2, \dots$$

Для задания переходной функции процесса выпишем цепочку результатов поступления на вход L сигналов x_t , $t = 1, 2, \dots$,

$$x_t \rightarrow x^t \rightarrow \zeta_t \rightarrow T_{\zeta_t, t} \rightarrow F_t (= T_{\zeta_t, t} F_{t-1}) \rightarrow y_t.$$

Если предположить, что правила F_t детерминированы, то $y_t = f(x^t, y^{t-1})$. Ниже мы рассматриваем для простоты этот случай.

Выпишем вероятность перехода от состояния множества C в подмножество $\tilde{S} \subseteq C$. Для этого выберем состояние в момент t

$$c' = (x^t; y^{t-1}; f_{t-1}),$$

а подмножество \tilde{S} изобразим в виде

$$\tilde{S} = (x^t M; y^{t-1}, (T_{\zeta_t, t} f_{t-1})(x^t, y^{t-1}); f_t),$$

где $M \in \mathfrak{M}$, $x^t M$ означает совокупность последовательностей x^{t+1} , в которой первые t элементов фиксированы (они образуют x^t), а x_{t+1} — произвольный элемент M . Для интересующей нас вероятности имеем (δ означает

символ Кронекера)

$$\mathcal{P}_{t+1}(c' \rightarrow \tilde{S}) = \mu_{t+1}(M | x^t; y^{t-1}, (T_{\zeta_{t_j}}, t f_{t-1})(x^t, y^{t-1})) \times \\ \times \delta_{f_t, T_{\zeta_{t_j}}, t f_{t-1}} \delta_{y, f_t}(x^t, y^{t-1}).$$

Построенный марковский процесс (C, \mathcal{P}) назовем *ассоциированным* с объектом $L \otimes \xi$. В рассмотренном здесь общем случае он неоднороден и транзитивен *) в усиленной форме: попав однажды в любое состояние, процесс никогда в дальнейшем в него не вернется. Для специальных классов управляемых процессов ассоциированный марковский процесс имеет более простую структуру.

Роль ассоциированных марковских процессов в теории адаптивных систем заключается в том, что они служат средством анализа обучаемых систем, с их помощью возможно устанавливать, является ли такая система адаптивной (для заданного класса $K \times \Phi$ и некоторой цели). В ряде случаев известные методы теории марковских процессов позволяют в принципе ответить на вопрос о достижении цели обучаемой системой и исследовать свойства адаптивной системы. Разумеется, и управляемый процесс, и обучаемая система должны быть достаточно простыми.

Рассмотрим обучаемые системы, представляющие собой автоматы. Пусть

$$A = (X, S, Y; \Pi_x, q),$$

где X и Y конечны, S не более чем счетно. Управляемые автоматом A процессы ξ_t имеют в качестве фазового пространства X , а в качестве пространства управлений Y . Сначала примем за ξ_t управляемую марковскую цепь с переходными вероятностями $\mu(x'' | x', y)$.

Множеством состояний ассоциированного марковского процесса служит $C = X \times S$, а вероятности переходов из $c' = (x', s_i)$ в $c'' = (x'', s_j)$ равны

$$p_{c'c''} = \sum_{y \in Y} \mu(x'' | x', y) \pi_{ij}(x'') q(y | s_i),$$

где $\pi_{ij}(x)$ — элемент матрицы Π_x вероятностей переходов автомата. Легко видеть, что матрица $\mathcal{P} = \|p_{c'c''}\|$ сто-

*) Это означает, что конечно математическое ожидание числа попаданий в каждое состояние.

хастическая. Если A — конечный автомат (множество S конечно), то множество C конечно и (C, \mathcal{P}) является ассоциированной марковской цепью. Задание на ξ_t функционала означает, что (C, \mathcal{P}) — однородная марковская цепь с доходами.

Укажем достаточное условие эргодичности цепи (C, \mathcal{P}) .

Если автомат A а) сильно связан *), б) не имеет циклических состояний **), а ξ_t — эргодическая цепь, то (C, \mathcal{P}) — эргодическая цепь. В самом деле, тогда все состояния C сообщающиеся и среди них нет циклических.

В предположении эргодичности цепи (C, \mathcal{P}) нетрудно указать выражение предельного математического ожидания дохода за 1 шаг. Предельное распределение порождается реализуемой автоматом A стратегией управления классом процессов $\{\xi_t\}$.

В частном случае, когда ξ_t есть ОПНЗ, цепь (C, \mathcal{P}) упрощается: тогда множество состояний есть $C=S$, а элементы матрицы $\mathcal{P}=\|p_{ij}\|$ равны

$$p_{ij} = P(s_i \rightarrow s_j) = \sum_{\substack{x \in X \\ y \in Y}} \mu(x | y) q(y | s_i) \pi_{ij}(x).$$

Снова цепь (C, \mathcal{P}) оказывается однородной. Легко убедиться, что если A — автомат с переменной структурой, свойство однородности нарушается.

В предположениях, что автомат A подчинен высказанным выше условиям, а управляемый им ОПНЗ ξ_t удовлетворяет условию

$$\mu(x | y) > 0$$

при всех x и y , ассоциированная марковская цепь (C, \mathcal{P}) эргодическая. Тогда существуют (и не зависят от начального состояния) предельные вероятности состояний π_j , $j=1, \dots, |S|$. С их помощью выписывается предельный

*) Сильная связность автомата означает, что любые его два состояния сообщающиеся (по крайней мере за $|S|$ шагов можно перейти из всякого состояния в другое).

**) Состояние s автомата называется *циклическим*, если у множества длин тех входных последовательностей, которые с положительной вероятностью переводят s в себя, общий наибольший делитель $d > 1$.

средний доход (мы имеем здесь в виду функционалы вида $\varphi_t = \varphi(x_t)$)

$$W(A, \xi) = \sum_{l=1}^k W(y_l) \sigma_l,$$

где $\sigma_l = \sum_j \pi_j q(y_l | s_j)$ — предельная вероятность действия y_l ,

а $W(y) = \int \varphi(x) \mu(dx | y)$ — математическое ожидание выигрыша за управление y . К этой величине сходится при неограниченном росте t математическое ожидание выигрыша автомата A спустя время t после начала управления процессом (начальным состоянием служит s)

$$W(A; s; t) = \sum_{l=1}^k W(y_l) \sum_j p_{s,j}^{(t-1)} q(y_l | p_j).$$

Иными словами, при любом s справедливо равенство

$$\lim_{t \rightarrow \infty} W(A; s; t) = W(A, \xi).$$

§ 4. Общие замечания об адаптивных системах

Мы рассматриваем класс K управляемых случайных процессов и класс Φ функционалов на его траекториях и пусть для любой пары $(\xi, \varphi) \in K \times \Phi$ задана достижимая (в принципе) цель управления. Предположим, что существует адаптивная система, обеспечивающая достижение этой цели на $K \times \Phi$. Спрашивается, позволяет ли это обстоятельство высказать какие-то требования к классам K , Φ и к цели. Иными словами, нас интересуют сейчас необходимые условия существования адаптивных систем. Важность решения этой проблемы очевидна, но без дополнительных ограничений на K , Φ и цель оно вряд ли существует. В самом деле, представим, что для реализуемой некоторой обучаемой системой L стратегии целью объявлены те фактические свойства, которые имеет функционал φ на траекториях процессов из K . При этом не исключается возможность того, что каждому процессу будет отвечать свое свойство функционала. Тогда L

окажется тривиальной адаптивной системой. Чтобы исключить такие патологические ситуации, неформально сузим рассматриваемые классы K и Φ .

1. Цель управления сформулирована в единообразном виде для всех пар (ξ, φ) , т. е. имеет стандартную математическую форму «максимизировать», «обеспечить неравенства» и т. д.

2. В множестве $K \times \Phi$ найдутся хотя бы две пары $(\xi^{(1)}, \varphi^{(1)})$, $(\xi^{(2)}, \varphi^{(2)})$, для которых приводящие к цели стратегии существенно различны.

3. Все меры $\mu_{t+1}(\cdot M | \xi^t, y^t)$ существенно зависят от управления y_t в момент t .

Смысл последнего условия состоит в том, что оно исключает неуправляемость процесса в некоторые моменты времени (или во все, начиная с некоторого) и вытекающую из этого возможность нарушения цели либо даже принципиальную ее недостижимость.

Накопленный опыт синтеза адаптивных систем и общие соображения дают основания полагать, что необходимые условия существования адаптивных систем заключаются в следующем:

I. Влияние давних действий на эволюцию процесса ослабевает с течением времени. В терминах управляемых условных вероятностей это записывается так:

$$\limsup_{\substack{s \rightarrow \infty \\ t > s + O(s)}} \mu_{t+1}(M | x^t, y_s^t \cdot y'^s) - \mu_{t+1}(M | x^t, y_s^t \cdot y''^s) = 0,$$

где y'^s , y''^s означают две последовательности управлений до момента s_0 . Это требование затухания можно относить к средним выигрышам, если цель управления формулируется с их помощью.

II. Существует приводящая к цели стационарная либо программная стратегия.

Пока мы не имеем строгого доказательства необходимости этих двух условий. Приведем поэтому эвристические соображения в их пользу. Нарушение первого условия влечет за собою то, что неправильно выбранные управление на начальном этапе сделают недостижимой цель управления. На раннем этапе функционирования адаптивная система наверняка совершает «неправильные» дей-

ствия. Более того, она руководствуется не теми правилами выбора действий. В этой ситуации поставленная цель, вообще говоря, недостижима. Рассмотрим для иллюстрации сказанного пример. Пусть цель управления относится к среднему выигрышу, вычисленному по мере, порожденной программной стратегией, т. е. к функции $W(y_0, y_1, \dots, y_t)$. Например, в качестве цели зададим максимизацию среднего выигрыша. Допустим, что множество управлений содержит два элемента $Y = \{y', y''\}$. Относительно класса управляемых случайных процессов K будем предполагать, что функции $W(y^t)$ имеют один знак (этот знак определяется процессом из класса и заранее неизвестен), если $y_0 = y'$, и другой знак в случае $y_0 = y''$. Ясно, что случайный выбор управления y_0 в начальный момент времени делает нереальным достижение цели для любого процесса из K .

Другое условие связано с тем, что если стратегия, которая приводит к цели, стационарная, то адаптивная система способна в ходе блуждания по множеству правил D приблизиться к правилу, порождающему стационарную стратегию. В противном случае, когда стратегию слагают более чем одно правило, сменяющие друг друга, адаптивная система должна разгадать закономерность чередования правил. Этого в условиях отсутствия полной информации об управляемом процессе ожидать трудно. Для всех классов процессов, для которых известны адаптивные системы управления, стационарные либо программные стратегии обеспечивают достижение цели.

Теперь обратимся к вопросу о достаточных условиях существования адаптивных систем. Подобного рода условий известно сейчас много: ими являются все теоремы о достижении обучаемой системой цели на классе управляемых случайных процессов. Мы рассмотрим здесь один общий принцип синтеза адаптивных систем, основанный на объединении методов математической статистики и теории управления.

Снова введем класс K управляемых случайных процессов, задаваемый семействами условных вероятностей (μ_t) . Пусть каждому такому семейству сопоставлен параметр q со значениями в евклидовом пространстве R , метрику в котором обозначим $\rho = \rho(q_1, q_2)$. Представим

себе, что параметр q допускает статистическую оценку по наблюдениям в ходе управления, для которой воспользуемся обычным обозначением $\hat{q}_t = \hat{q}(\xi^t, y^t)$, причем эта оценка является состоятельной:

$$\lim_{t \rightarrow \infty} \rho(q, \hat{q}_t) = 0,$$

а предел понимается в каком-нибудь из распространенных теоретико-вероятностных смыслов (по вероятности, с вероятностью единица, в среднем квадратическом и пр.). Семейства условных вероятностей, отвечающие значению \hat{q}_t параметра, принадлежат по принимаемому условию к классу K . Допустим, что известен алгоритм построения «оптимальной» стратегии $\sigma(q)$, приводящей к цели процесс из K , со значением параметра q . Конкретизируем цель — максимизировать средний выигрыш $W(t)$ — математическое ожидание функционала φ_t на траекториях управляемого процесса. В обозначениях § 4 гл. I требуется, чтобы для любого $\epsilon > 0$ при всех $t > t_\epsilon$ выполнялось неравенство $W_t > \bar{W} - \epsilon$ (асимптотическая оптимальность в слабом смысле). Выделим с помощью параметра q процесс из K и через $\tilde{\sigma}(q)$ обозначим такую стратегию для него: в момент t по очередному значению \hat{q}_t оценки параметра вычисляются правила выбора действий, входящие в «квазиоптимальную» стратегию $\tilde{\sigma} = \sigma(\hat{q}_t)$. Эти правила используются до следующего момента получения оценки. Сделаем последнее предположение: математическое ожидание $E\varphi_t = W(t)$, найденное по мере, порожденной стратегией $\tilde{\sigma}$, стремится к \bar{W} с ростом t .

Теперь становится очевидной структура адаптивной системы для класса K . Она параллельно осуществляет две функции: оценивает параметр q управляемого ею процесса и уточняет правила выбора управления. В этой обучаемой системе правила трансформируются операторами T (здесь это алгоритм синтеза оптимальной стратегии), зависящими от статистики \hat{q}_t , означающей текущую оценку параметра q . В сделанных предположениях эта система гарантирует асимптотическую оптимальность в слабом смысле.

Изложенный здесь общий принцип синтеза адаптивных систем допускает разного рода модификации. Так иногда

естественно предполагать, что «квазиоптимальная» стратегия $\tilde{\sigma}(q)$ «сходится» к оптимальной стратегии $\sigma(q)$. Это означает, что в подходящей метрике справедливо равенство

$$\lim_{t \rightarrow \infty} [\tilde{F}_t - F_t] = 0,$$

где \tilde{F}_t и F_t — соответственно правила выбора управления, назначаемые стратегиями $\tilde{\sigma}$ и σ . В таких случаях приходится доказывать сходимость $\tilde{W}(t) \rightarrow W$. Нетрудно представить себе иные формы этого метода, который называют «идентификационным». Ранее (во введении) мы уже отмечали, что метод не универсален хотя бы по двум причинам: не всегда существуют состоятельные оценки параметра и не всегда известен алгоритм построения оптимальной стратегии. Рассмотрим один подход к задаче управления марковскими процессами, который иногда считают аддитивным.

Рассматривается класс Q управляемых марковских процессов ξ_t с евклидовыми пространствами X и Y , задаваемый переходными функциями $\mu(M | \xi, y; q)$, где q означает параметр, характеризующий процесс в классе Q . Пусть этот процесс наблюдаемый. В каждый момент времени определен доход $\varphi_t = \varphi_t(\xi_t, y_{t-1})$ — непрерывная ограниченная функция на $X \times Y$. Зафиксирован отрезок времени $[1, T]$ и требуется максимизировать математическое ожидание величины

$$J = \sum_{n=1}^T \varphi_n(\xi_n, y_{n-1}),$$

т. е. найти «оптимальную» стратегию, приводящую к максимуму функции

$$R = EJ = \sum_{n=1}^T E \varphi_n(\xi_n, y_{n-1}).$$

В предположении, что известна переходная функция процесса из класса Q , эта экстремальная задача решается классическими методами. Обратимся к ее «аддитивному» аналогу. Теперь мы считаем, что параметр q неизвестен, т. е. задано множество переходных функций $\{\mu(\cdot | \cdot \dots ; q), q \in Q\}$ и наблюдаема траектория управляемого процесса.

Требуется найти стратегию, которая максимизирует функцию R , т. е. найти оптимальную совокупность рандомизированных правил выбора действий F_0, F_1, \dots, F_{T-1} .

Воспользуемся байесовским подходом к задачам с неизвестным параметром и примем гипотезу, что существует и известна априорная плотность вероятностей $p_0(q)$. Тогда правило Байеса позволяет по результатам наблюдений за процессом находить апостериорные плотности параметра q .

Займемся построением стратегии, основываясь на идеях динамического программирования. Будем считать, что встречающиеся ниже распределения имеют плотности, которые мы станем отличать друг от друга по обозначениям аргументов.

Начнем с отыскания последнего (по времени) правила $p_{T-1}(y_{T-1} | \xi^{T-1})$ плотности распределения вероятностей, которое задает управление y_{T-1} . Оно находится из принятого нами условия, что достигается максимум функционала

$$\mathbf{E}(\varphi_T | x^{T-1}) = \int \varphi_T(x_T, y_{T-1}) p(x_T, y_{T-1} | x^{T-1}) dx_T dy_{T-1}$$

для любой предыстории x^{T-1} . Выпишем фигурирующую в этом интеграле плотность

$$\begin{aligned} p(x_T, y_{T-1} | x^{T-1}) &= p(y_{T-1} | x^{T-1}) p(x_T | y_{T-1}, x^{T-1}) = \\ &= p(y_{T-1} | x^{T-1}) \int p(x_T | x_{T-1}, y_{T-1}; q) p(q | x^{T-1}) dq, \end{aligned}$$

вторая плотность в интеграле справа вычисляется по формуле Байеса. Максимизируемый функционал запишем в виде

$$\mathbf{E}(\varphi_T | x^{T-1}) = \int \varphi_T(y_{T-1}, x^{T-1}) p(y_{T-1} | x^{T-1}) dy_{T-1},$$

где принято обозначение

$$\begin{aligned} \psi_T(y_{T-1}, x^{T-1}) &= \\ &= \int \varphi_T(x_T, y_{T-1}) p(x_T | x_{T-1}, y_{T-1}; q) p(q | x^{T-2}) dx_T dq. \end{aligned}$$

Пусть y_{T-1}^* — значение аргумента y_{T-1} , при котором ψ_T достигает максимума. Это значение зависит от набора ве-

личин x^{T-1} , т. е. можно записать $y_{T-1}^* = y_{T-1}^*(x_0, x_1, \dots, x_{T-1})$. Сделанное допущение о существовании максимума ψ_T справедливо при надлежащих ограничениях на плотности, которые находятся под знаком интеграла (определяющего функцию ψ_T).

Ясно, что плотность $p(y_{T-1} | x^{T-1})$ должна быть сосредоточена в точке y_{T-1}^* , т. е. представлять собой « δ -функцию» $p^*(y_{T-1} | x^{T-1}) = \delta(y_{T-1} - y_{T-1}^*)$.

Попытаемся найти теперь следующее правило F_{T-2} с плотностью $p(y_{T-2} | x^{T-2})$ из условия, чтобы вместе с найденным правилом $p^*(y_{T-1})$ они доставили максимум функционалу $E(\varphi_T + \varphi_{T-1})$ или, что то же самое, функционалу $E(\varphi_T + \varphi_{T-1} | x^{T-2})$ при любых предысториях x^{T-2} . Аналогично приведенному выше рассуждению имеем

$$E(\varphi_{T-1} | x^{T-2}) = \int \varphi_{T-1}(y_{T-2}, x^{T-2}) p(y_{T-2} | x^{T-2}) dy_{T-2},$$

где принято обозначение

$$\begin{aligned} \varphi_{T-1} = & \int \varphi_{T-1}(x_{T-1}, y_{T-2}) p(x_{T-1} | x_{T-2}, y_{T-2}; q) \times \\ & \times p(q | x^{T-2}) dx_{T-1} dy_{T-2} dq. \end{aligned}$$

Допустим, что вторая плотность под интегралом вычислена. В силу равенств

$$E(\varphi_T | x^{T-2}) = E[E(\varphi_T | x^{T-1}) | x^{T-2}],$$

$$E(V_T | x^{T-2}) = \int V_T p(x_{T-1} | x_{T-2}, y_{T-2}) p(y_{T-2} | x^{T-3}) dy_{T-2} dx_{T-1},$$

где $V_T = \max \varphi_T$, имеем

$$\max \varphi_{T-1} + \max E(\varphi_T | x^{T-2}) =$$

$$= \max \int [\psi_{T-1} + \int V_T p(x_{T-1} | x_{T-2}, y_{T-2})] p(y_{T-2} | x^{T-2}) dy_{T-2}.$$

Правило выбора действия в момент $T-2$ находим, как и ранее, в виде функции y_{T-2}^* от предыстории x^{T-2} процесса. Это такая функция, которая максимизирует стоящее под

интегралом справа выражение в квадратных скобках. Иными словами, искомое правило снова нерандомизированное — порождающая его плотность является δ -функцией $p^*(y_{T-2} | x^{T-2}) = \delta(y_{T-2} - y_{T-2}^*)$, сосредоточенной в точке y_{T-2}^* . Поступая аналогичным образом, шаг за шагом мы найдем правила, слагающие оптимальную стратегию. В сделанных предположениях эта стратегия нерандомизированная. Обратим внимание на то, что правила выбора управлений находятся с конца, а необходимые для этого апостериорные плотности распределения параметра $p(q | x^t)$ — с начала. Изложенный вывод носит эвристический характер и не имеет, конечно, доказательной силы. Однако в тех случаях, когда эти рассуждения справедливы, они проливают свет на структуру оптимальной стратегии. Заметим, что можно было бы рассматривать более сложную ситуацию — наблюдается не процесс ξ_t , а некоторая его функция, к тому же искаженная помехами. Например, наблюдают процесс $\eta_t = h(\xi_t, \zeta_t)$, где ζ_t — последовательность независимых случайных величин, распределение которых неизвестно (его снова можно представить себе зависящим от неизвестного параметра).

Обратимся к обсуждению изложенной здесь процедуры решения оптимизационной задачи. Прежде всего, допущение о существовании априорного распределения параметра q , подразумевающее его стохастическую природу, не эквивалентно отсутствию сведений о значении q для конкретного управляемого процесса. Естественно, что произвольное задание $p(q)$ не имеет «rationального» основания и, что является весьма важным, меняет цель управления. В самом деле, мы вначале потребовали максимизации функционала $R = EJ$, где математическое ожидание берется по мерам, порожденным допустимыми стратегиями. Принятие байесовской концепции означает, что произведено еще одно интегрирование, причем соответствующая мера не имеет никакого отношения к рассматриваемой ситуации (виду экстремальной задачи, классу управляемых процессов). Эта мера является «априорным» распределением. В результате оказывается, что указанная выше «оптимальная» стратегия не обеспечивает величину максимума R , достижимого в первоначальной постановке,

а меньшую величину, зависящую от наименее назначенного априорного распределения. В таких условиях нельзя говорить об ϵ -оптимальности полученной стратегии, потому что априорное распределение не выбирается сосредоточенным «вокруг» истинного значения параметра. Более того, отсутствие сведений о конкретном процессе препятствует такому выбору распределения, вынуждая делать его «размазанным» по всему пространству Q . Таким образом, изложенная здесь комбинация динамического программирования (вычисляющего стратегию) и байесовского подхода (оценивающего свойства процесса) не является адаптивной системой, как не обеспечивающая поставленной цели управления.

Естественно увязать недостижимость цели рассмотренным алгоритмом с конечностью промежутка времени, к которому относится экстремальная задача. Если цель относится к функционалу на всей (бесконечной) траектории процесса, например, к величине

$$R = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{n=1}^T E_{\varphi_n}(\xi_n, y_{n-1}),$$

то байесовский подход дает в пределе истинное значение параметра. В самом деле, известно, что при очень широких допущениях апостериорные распределения сходятся к « δ -функции», сосредоточенной в точке q . Тогда допущенные в начальной стадии управления ошибки окажутся со временем нивелированными и будет возможно достичь максимума (либо с точностью до ϵ верхней грани) функции R . Следует, однако, заметить, что в задаче с бесконечным временем примененный выше метод динамического программирования становится непригодным. В последующих главах излагаются конструкции адаптивных систем для задач рассмотренного здесь типа.

ГЛАВА III

АВТОМАТЫ ДЛЯ УПРАВЛЕНИЯ ОДНОРОДНЫМИ ПРОЦЕССАМИ С НЕЗАВИСИМЫМИ ЗНАЧЕНИЯМИ

§ 1. Постановка задачи

Через ξ_t обозначим ОПНЗ, задаваемый, как обычно, управляемой условной вероятностью $\mu(M|y)$, $M \in \mathfrak{M}$, $y \in Y$. Будем всюду в этой главе считать конечным пространство управлений $Y = \{y_1, \dots, y_k\}$ и, если не оговорено противное, фазовое пространство $X = \{x_1, \dots, x_m\}$. Ясно, что условная вероятность $\mu(M|y)$ удовлетворяет требованиям из § 3 гл. I. Можно рассматривать вероятности отдельных значений $x \in X$ при данном управлении $y \in Y$ — $\mu(x|y)$. Нумеруя элементы $x \in X$ и $y \in Y$, приходим к $(m \times k)$ -матрице $\mu = [\mu_{ij}]$, где $\mu_{ij} = \mu(x_i | y_j)$.

Введем функционал $\varphi_t = \varphi(\xi_t)$, который также представляет собой ОПНЗ. К этому функционалу будет отнесена далее цель управления процессом. Если исходный процесс наблюдаем, то и значения φ_t наблюдаемы (они заведомо могут вычисляться в управляющей системе). Если же ξ_t не наблюдаем, то следует предположить, что наблюдаемы φ_t или его взаимно однозначная функция. В таком случае естественно отождествить наблюдаемый и управляемый процессы. Поэтому множество X мы далее примем числовым и значения процесса трактуем как доходы.

Средний выигрыш (доход) в момент t равен

$$W(y) = E\xi_t = \sum_{i=1}^m x_i \mu(x_i | y),$$

$$W(y) = E\xi_t = \sum_{i=1}^m x_i \mu(x_i | y),$$

где y — действие, совершенное в момент $t-1$. При конечном Y средние выигрыши образуют совокупность из k

чисел; максимальное из них $\bar{W} = \max_y W(y)$. В общем случае их недостаточно для того, чтобы однозначно восстановить вероятности $\mu(x|y)$. Действительно, последних mk (независимых $(m-1)k$), а средних выигрыш всего k . В одном случае числа $W(y)$ полностью определяют процесс. Это бинарные ОПНЗ, у которых фазовое пространство содержит всего два элемента $\bar{X} = (x_1, x_2)$. Сопоставим им числа: $x_1 \rightarrow 1$, $x_2 \rightarrow -1$. Условную вероятность x_1 при управлении y_i обозначим q_i , а вероятность x_2 обозначим $p_i = 1 - q_i$. Тогда $W(y_i) = q_i - p_i$, $i = 1, \dots, k$. Отсюда и из равенства $p_i + q_i = 1$ следует

$$q_i = \frac{1 + W(y_i)}{2}, \quad p_i = \frac{1 - W(y_i)}{2}, \quad i = 1, \dots, k.$$

Мы рассматриваем класс K всех ОПНЗ с одинаковыми пространствами X и Y . Обучаемые системы для них будем строить в виде автоматов с конечным или счетным множеством состояний. Как мы знаем, определено математическое ожидание величины ξ_t в момент t , если управляющий автомат A находился в начальный момент в состоянии s . Сформулируем в качестве цели управления ε -оптимальность в слабом смысле, т. е. выполнение неравенства ($\varepsilon > 0$ фиксировано) (см. § 3 гл. II)

$$W(A; s, t) > \bar{W} - \varepsilon$$

при всех $t > t_\varepsilon$ и любом начальном состоянии автомата.

Высказанная цель управления означает, что следует построить семейство автоматов $\{A\}$ таких, что для каждого ε существуют автоматы из семейства, которые обеспечивают выполнение требуемых неравенств.

Забегая вперед, отметим, что автоматы реализуют нестационарную стратегию. Кроме того, в ходе управления они не решают задачу оценки характеристик ОПНЗ. Применительно к рассматриваемой нами задаче смысл слова «оценка» состоит в построении k условных распределений $\mu(M|y)$, $y \in Y$. Этой проблемой не занимается ни один автомат. Принцип функционирования всех этих автоматов заключается в тем более частом совершении действия, чем оно «выгоднее», т. е. чём выше отвечающий ему средний выигрыш. Этот принцип влечет за собой

ϵ -оптимальность в сильном смысле. Для доказательства утверждения введем обозначения: $N_i(t)$ — количество совершений действия y_i за первые t тактов функционирования автоматов, $V_i(t)$ — эмпирический средний выигрыш автомата, полученный за y_i в течение первых t тактов. Эта величина равна по определению

$$V_i(t) = \frac{1}{N_i(t)} \sum_{l=1}^{N_i(t)} x_l^{(i)},$$

где $(x_l^{(i)})$ — реализация ОПНЗ после выбора управления y_i . Ясно, что если $\lim_{t \rightarrow \infty} N_i(t) = \infty$, то с вероятностью единица

$$\lim_{t \rightarrow \infty} V_i t = W(y_i).$$

Предположим, что ассоциированная с $A \otimes \xi$ марковская цепь эргодическая. Тогда с вероятностью единица

$$\lim_{t \rightarrow \infty} \frac{N_i(t)}{t} = \sigma_i, \quad i = 1, \dots, k,$$

где σ_i — предельные вероятности совершения действий y_i . С их помощью выражается предельный средний выигрыш автомата

$$W(A) = \sum_{i=1}^k W(y_i) \sigma_i.$$

По свойству ϵ -оптимальности в слабом смысле следующие соотношения:

$$\lim_{t \rightarrow \infty} W(A; s, t) = W(A) > \bar{W} - \epsilon$$

справедливы для любого ОПНЗ (с одинаковыми X и Y).

Образуем эмпирический суммарный выигрыш автомата A за время t

$$V(t) = \frac{1}{t} \sum_{l=1}^t x_l.$$

Запишем его в таком виде:

$$V(t) = \sum_{j=1}^k \frac{N_j(t)}{t} V_j(t).$$

Из сказанного ранее следует, что с вероятностью единица

$$\lim_{t \rightarrow \infty} V(t) = \sum_{j=1}^k W(y_j) \sigma_j.$$

Это означает, что начиная с некоторого (случайного) момента времени τ выполняется неравенство

$$V(t) > \bar{W} - \epsilon,$$

причем $P(\tau_\epsilon < \infty) = 1$. Это доказывает, что рассматриваемые нами автоматы, преследующие в качестве цели ϵ -оптимальность в слабом смысле, достигают также ϵ -оптимальности в сильном смысле.

Нас будет интересовать возможность использования рассматриваемых автоматов для управления процессами с бесконечными пространствами X , в частности, представляющими собой отрезок числовой прямой. Окажется, что исследуемые конструкции допускают такое расширение.

§ 2. Конечные автоматы для бинарных ОПНЗ

Всюду в этом параграфе ξ_i означает бинарный ОПНЗ. Наглядности ради иногда будем называть его значение x_1 (к которому сопоставлено число 1) «поощрением», а x_2 (ему отвечает число -1) — «наказанием». Мы уже знаем, что такой процесс полностью характеризуем средними выигрышами — числами $W_i = W(y_i)$, $i = 1, \dots, k$. Мы строим здесь ϵ -оптимальные семейства конечных автоматов, обеспечивающих ϵ -оптимальность в сильном смысле. Прежде всего надлежит задать параметр, определяющий семейство автоматов, а потом выбрать структуру автоматов.

Отметим, что семейство вероятностных автоматов с ограниченным сверху числом состояний не может быть ϵ -оптимальным семейством в классе всех бинарных ОПНЗ. В качестве индекса автомата A такого семейства примем

число состояний N , т. е. семейство записывается в виде (A_N) .

Рассматриваются автоматы Мура $A_N = (X, S_N, Y; \Pi_x^{(N)}, q_N)$, где $X = (x_1, x_2)$, $Y = (y_1, \dots, y_k)$, $\Pi_x^{(N)}$ — функция переходов, задаваемая при каждом N двумя стохастическими матрицами $\Pi_{x_1}^{(N)}$, $\Pi_{x_2}^{(N)}$ вероятностей переходов, и q_N — функция выходов, однозначное отображение $S_N \rightarrow Y$. Отсюда по муровскому свойству вытекает, что множество состояний S_N разлагается в прямую сумму (объединение непересекающихся подмножеств):

$$S_N = S_1^{(N)} + \dots + S_k^{(N)},$$

где $S_j^{(N)}$ — множество всех тех и только тех состояний, которым отвечает выходной сигнал y_j . Сопоставим автомatu A_N подавтоматы $A_{N,j} = (X, S_j^{(N)}, y_j, \Pi_x^{(N,j)})$, $j = 1, \dots, k$, с единственным выходным сигналом y_j и функцией переходов $\Pi_x^{(N,j)}$, индуцированной на подмножестве $S_j^{(N)}$ функцией переходов $\Pi_x^{(N)}$.

Сформулируем допущения о структуре автоматов ε -оптимальных семейств.

А. Из любого подмножества $S_j^{(N)}$ возможен переход в любое другое подмножество.

Условимся рассматривать лишь следующие два варианта переходов от состояний одного подавтомата к состояниям другого: 1) циклический — от $S_j^{(N)}$ к $S_{j+1}^{(N)}$ (символически $S_j^{(N)} \rightarrow S_{j+1}^{(N)}$, причем $S_k^{(N)} \rightarrow S_1^{(N)}$). Это детерминированный перебор всех действий подряд; 2) равновероятный — от любого подавтомата переход в остальные подавтоматы совершается с одинаковыми вероятностями (равными $1/k$), причем разрешается возврат в исходный подавтомат.

Отметим в каждом подавтомате входные состояния, в которых оказывается автомат, когда он из состояний других подавтоматов попадает в состояния данного подавтомата, и выходные состояния, в них автомат находится перед тем, как на следующем такте покинуть данный подавтомат. Назовем автомат одновходовым (одно-

выходовым), если все подавтоматы $A_1^{(N)}, \dots, A_k^{(N)}$ имеют по единственному входному (выходному) состоянию. В противном случае говорят о многовходовых (многовыходовых) автоматах.

В. Все подавтоматы $A_1^{(N)}, \dots, A_k^{(N)}$ изоморфны (т. е. отличаются лишь обозначением состояний).

Отсюда вытекает, что подмножества состояний $S_j^{(N)}$ имеют одинаковое число элементов. Обозначим его n (тогда $N=kn$) и будем называть это число «памятью» автомата. Теперь множество $S_j^{(N)}$ обозначаем $S_j^{(n)}$, а его элементы $s_i^j, i=1, \dots, n$, с выполнением условия, что взаимно соответствующие элементы разных множеств получают одинаковые нижние индексы. Взаимно соответствующими являются, в частности, входные и выходные состояния.

С. Каждый подавтомат $A_j^{(N)}$ сильно связан, т. е. под действием надлежащей последовательности входных сигналов (длиною, не превосходящей n) положительна вероятность перейти из любого состояния подавтомата в любое, и не содержит циклических состояний.

Из высказанных предположений следует, что автоматы $A^{(N)}$ ε -оптимального семейства сильно связаны и не содержат циклических состояний.

Этот вывод приводит к важному заключению относительно ассоциированной марковской цепи $(S^{(N)}, \mathcal{P}^{(N)})$: эта цепь эргодическая.

Математическое ожидание выигрыша в момент времени t автомата $A^{(N)}$, находившегося в начальный момент в состоянии s_0 , равно

$$W(A^{(N)}; s_0, t) = \sum_{i=1}^k W_i P_{s_0, i}^{(N)}(t-1),$$

где $P_{s_0, i}^{(N)}(t-1) = \sum_{l \in S_i} p_{s_0, l}^{(N)}(t-1)$, а $p_{s_0, l}^{(N)}(t-1)$ — элементы матрицы $\mathcal{P}^{(N)^{t-1}}$ (здесь используется единая нумерация всех состояний автомата). В силу эргодичности марковской цепи существует и не зависит от начального состояния предел (стремление к пределу происходит с экспоненциальной скоростью)

$$\lim_{t \rightarrow \infty} P_{s_0, i}^{(N)}(t) = \sigma_i^{(N)}.$$

Поэтому предельный средний выигрыш автомата $A^{(N)}$ равен

$$W(A^{(N)}) = \sum_{i=1}^k W_i \sigma_i^{(N)}.$$

Требование ϵ -оптимальности в слабом смысле означает, что при достаточно большом числе состояний n в каждом из подмножеств $S_1^{(N)}, \dots, S_k^{(N)}$ (здесь и далее k фиксировано) должно выполняться неравенство $W(A^{(kn)}) > \bar{W} - \epsilon_n$. Из него следует, что предельная вероятность оптимального действия (или сумма этих вероятностей, если оптимальных действий несколько) должна стремиться к 1 при неограниченном росте n . Иными словами, справедливо следующее утверждение.

Теорема 1. Для ϵ -оптимальности в слабом смысле семейства конечных автоматов $A^{(kn)}$, управляемых бинарными ОПНЗ, необходимо и достаточно, чтобы

$$\lim_{n \rightarrow \infty} \frac{\sigma_j^{(kn)}}{\sum_{i \in I} \sigma_i^{(kn)}} = 0, \quad j \notin I,$$

где I — совокупность индексов оптимальных действий, а $\sigma_j^{(kn)}$ — предельная вероятность действия y_j .

Основываясь на этом результате, для установления того, является ли некоторое семейство $A^{(kn)}$ конечных автоматов ϵ -оптимальным или нет, необходимо вычислить предельные вероятности состояний ассоциированной марковской цепи (S, \mathcal{P}) . Эта процедура, всегда допустимая для эргодических цепей, может быть весьма трудоемкой, и поэтому желательно установить связи предельных вероятностей действий с иными характеристиками автоматов, вычисление которых не требует решения системы линейных алгебраических уравнений.

Изучаемые нами семейства конечных автоматов, управляемые бинарными ОПНЗ, тем дольше пребывают в подмножествах состояний $S_1^{(kn)}, \dots, S_k^{(kn)}$, чем чаще в ответ на каждое действие поступает поощрение (т. е. чем больше средний выигрыш). Причиной покидания такого подмножества служит появление на входе автомата достаточно

длинной цепочки наказаний. Поэтому естественно ввести такую характеристику автоматов — среднее время пребывания в подмножестве $S_j^{(kn)}$ (или совершения подряд действия y_j), которое обозначим

$$T_j^{(kn)} = T_j^{(kn)}(W_1, \dots, W_k).$$

Вычисляется эта характеристика обычно методами теории случайных блужданий.

Заметим, что понятие среднего времени пребывания в подмножестве состояний $S_j^{(kn)}$ имеет ясный смысл для одновходовых и одновыходовых автоматов. В противном случае определение этого понятия наталкивается на трудности.

Установим связь предельных вероятностей действия автоматов $A^{(kn)}$ со средними временами T_j , опуская пока индексы n .

Теорема 2. Пусть A — конечный одновходовый и одновыходовый автомат Мура с детерминированной функцией выхода, с циклическим либо равновероятным правилом смены действий. Пусть такой автомат управляет бинарным ОПНЗ и ассоциированная марковская цепь эргодическая. Тогда при любых $i, j=1, \dots, k$

$$\frac{\sigma_i}{\sigma_j} = \frac{T_i}{T_j}.$$

Доказательство сначала проведем для автоматов с циклической сменой действий. Имеем

$$\sigma_i = \lim_{T \rightarrow \infty} \frac{\mathbb{E} \sum_{t=1}^T \zeta_t^{(i)}}{T},$$

где

$$\zeta_t^{(i)} = \begin{cases} 1, & \text{если } s(t) \in S_i, \\ 0, & \text{если } s(t) \notin S_i. \end{cases}$$

Отсюда находим

$$\frac{\sigma_i}{\sigma_j} = \lim_{L \rightarrow \infty} \frac{E \sum_{t=1}^L \zeta_t^{(i)}}{E \sum_{t=1}^L \zeta_t^{(j)}} = \lim_{L \rightarrow \infty} \frac{E \sum_{l=1}^{\mu_L^i} \tau_l^i}{E \sum_{l=1}^{\mu_L^j} \tau_l^j}. \quad (1)$$

В правой части равенства фигурируют случайные величины τ_l^i (τ_l^j), означающие длины промежутков времени, проведенные автоматом в состояниях из S_i (S_j). Имеем, по определению, $E\tau_l^i = T_i$, $E\tau_l^j = T_j$, $l = 1, 2, \dots, \mu_L$. Целочисленные случайные величины μ_L^i и μ_L^j указывают, сколько раз за время L автомат попадал в группы состояний S_i и S_j . Ясно, что $\lim_{L \rightarrow \infty} \mu_L^i = \infty$ и в силу способа смены действий $|\mu_L^i - \mu_L^j| \leq 1$. Воспользовавшись тождеством Вальда

$$E \sum_{l=1}^{\mu_L^i} \tau_l^i = T_i E \mu_L^i$$

(такое же соотношение выполняется для знаменателя) и равенством (1), приходим к утверждению теоремы.

В случае равновероятного перехода от действия к действию легко проверяется, что выполняется равенство

$$\lim_{L \rightarrow \infty} \frac{E \mu_L^j}{E \mu_L^i} = 1,$$

которое вместе с тождеством Вальда и (1) доказывает теорему.

Ниже воспользуемся такими следствиями теоремы 2:
Следствие 1. Предельные вероятности действий и средние времена связаны равенствами

$$\sigma_j = \frac{T_j}{\sum_{l=1}^k T_l}, \quad j = 1, \dots, k.$$

Следствие 2. Предельный средний выигрыш автомата A при управлении ОПНЗ равен

$$W(A) = \frac{\sum_{i=1}^k W_i T_i}{\sum_{i=1}^k T_i}.$$

Отметим, что в предположении одновходовости автомата среднее время $T_j = T_j(W_j)$ зависит лишь от среднего выигрыша W_j . Из изоморфизма подавтоматов A_j вытекает, что все эти функции одинаковы, т. е.

$$T_j = T(W_j).$$

Этот факт облегчает исследование автоматов определенного здесь типа, так как требует вычисления лишь одной функции $T(W)$. Дальнейшего упрощения мы добьемся, если будем считать граф подавтомата A_j линейным (т. е. представляющим собою последовательность точек на прямой).

Условимся снабжать автоматы $A^{(n)}$ из ϵ -оптимального семейства двумя индексами. Будем писать $A_{k,n}$, где k означает число действий автомата, а n — количество состояний в каждом подавтомате.

Выражение предельных средних выигрышей изучаемых автоматов (по следствию 2 теоремы 2) таково:

$$W(A_{k,n}) = \frac{\sum_1^k W_i T(W_i)}{\sum_1^k T(W_i)}.$$

Сопоставление этого равенства с теоремой 1 приводит к следующему утверждению.

Теорема 3. Для ϵ -оптимальности в слабом смысле в классе бинарных ОПНЗ семейства автоматов $A_{k,n}$, удовлетворяющих условиям теоремы 2, необходимо и достаточно

$$\lim_{n \rightarrow \infty} \frac{T^{(n)}(W')}{T^{(n)}(W'')} = 0$$

при любых W' и W'' , связанных неравенством $W'' > W'$.

Перейдем к конкретным примерам ϵ -оптимальных семейств автоматов.

Автоматы $D_{k,n}$ («глубокие») определяются графом на рис. 3. Каждой ветви состояний s_i^j , $j=1, \dots, k$; $i=1, \dots, n$, отвечает одно и то же действие y_j . Состояние s_i^j является единственным входным и выходным. Поощрения переводят автомат из состояния s_i^j в s_n^j , а наказания — из s_i^j

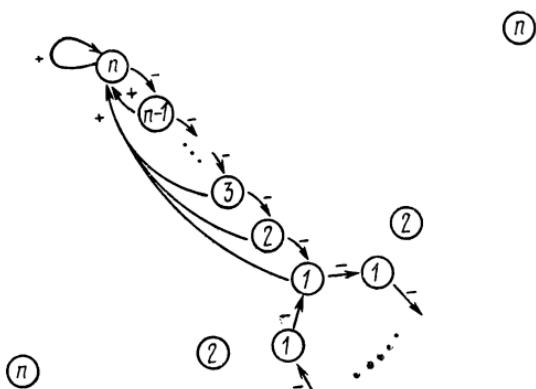


Рис. 3.

в s_{i+1}^j , причем из s_i^j совершается переход либо в s_{i+1}^{j+1} (при циклической смене действий), либо в любое входное состояние $s_1^1, s_1^2, \dots, s_1^k$ с равными вероятностями. Изоморфизм всех ветвей (или, что то же самое, подавтоматов) позволяет исследовать среднее время пребывания лишь на одной из них. Воспользуемся с этой целью методами теории случайного блуждания.

Обозначим через T_x среднее время пребывания на ветви при условии, что начальным состоянием является s_x , $x=1, \dots, n$. Интересующая нас величина $T^{(n)}(W)$ совпадает с T_1 , если движение вправо (в состояние s_n) происходит с вероятностью $q = \frac{1+W}{2}$, а влево — на один шаг — с вероятностью $p = \frac{1-W}{2}$. Нетрудно убедиться, что T_x удовлетворяет разностному уравнению

$$T_x = pT_{x-1} + qT_n + 1, \quad x=1, \dots, n, \quad (2)$$

при граничном условии $T_0=0$. Заметим прежде всего, что

$$T_1 = qT_n + 1,$$

и далее получаем

$$T_2 = pT_1 + qT_n + 1 = pT_1 + T_1 = T_1(1 + p).$$

Очевидно, при всех $x=2, \dots, n$ имеем

$$T_x = T_1 \sum_{l=p}^{x-1} p^l = T_1 \frac{1-p^x}{q}.$$

В частности, при $x=n$, $T_n = T_1 \frac{1-p^n}{q}$. Сравнивая это с выражением для T_1 , находим

$$T_1 = qT_1 \frac{1-p^n}{q} + 1 = T_1(1 - p^n) + 1.$$

Это приводит нас к искомому среднему времени

$$T^{(n)}(W) = \frac{1}{p^n} = \left(\frac{2}{1-W}\right)^n.$$

Если $W=1$, то $q=1$ и автомат $D_{k,n}$, оказавшись на соответствующей ветви, никогда ее не покинет. Если же $W=-1$, то $q=0$ и на соответствующей ветви автомат получает только наказания и может находиться на ней только один такт. При всех прочих значениях $W \in (-1, 1)$ среднее время пребывания на ветви растет экспоненциально с увеличением числа состояний n .

Из выражения $T^{(n)}(W)$ видно, что условие теоремы 3 выполнено и автоматы $D_{k,n}$ образуют ϵ -оптимальное семейство в классе всех бинарных ОПНЗ. Предельный средний выигрыш равен

$$W(n) = W(D_{k,n}) = \frac{\sum_{i=1}^k \frac{W_i}{(1-W_i)^n}}{\sum_{i=1}^k \frac{1}{(1-W_i)^n}}.$$

Пусть максимальным средним выигрышем служит W_1 (существование нескольких максимальных среди W_i не меняет последующих заключений). Оценим скорость сходимости $W(n)$ к W_1 :

$$W_1 - W(n) = \frac{\sum_{i=2}^k (W_1 - W_i) h(W_{i,n})}{1 + \sum_{i=2}^k h(W_{i,n})},$$

$$h(W_{j,n}) = \left(\frac{1-W_1}{1-W_j}\right)^n.$$

Эта разность положительна (если, разумеется, не все W_i равны W_1). Введем обозначения: $\mu = \max_{i \neq 1} \frac{1-W_1}{1-W_i}$, $c = \max_j (W_1 - W_j)$. Тогда

$$W_1 - W(n) \leq c\mu^n = ce^{-\lambda n}, \quad \lambda = \ln \mu^{-1} > 0.$$

Следовательно, с ростом числа состояний n предельный средний выигрыш экспоненциально быстро стремится к максимуму W_1 . Полученные здесь результаты приводят к выводу о равномерной сходимости средних выигрышей $W(D_{k,n})$ на множестве всех бинарных ОПНЗ. Это сразу следует из теоремы Дини *).

Перейдем к другому примеру. Охарактеризуем сначала класс семейства автоматов $Q_{k,n}$. Их графы снова имеют звездчатый вид, а ветви линейно упорядочены (рис. 4). Эти автоматы называют *квазилинейными*. Переходы осуществляются следующим образом. При поощрениях возможны два перехода из состояния s_i : с вероятностью q_+ в s_{i+1} и с вероятностью $p_+ = 1 - q_+$ в s_{i-1} . Поощрение в состоянии s_1 выводит с вероятностью p_+ из рассматриваемой ветви. Наказание влечет такие же переходы, но с иными вероятностями: из s_i в s_{i-1} с вероятностью p_- (из s_1 возможен выход из ветви), из s_i в s_{i+1} с вероят-

*) Если монотонно возрастающая последовательность непрерывных функций сходится на конечном отрезке к непрерывной функции, то последовательность сходится равномерно.

ностью $q_- = 1 - p_-$ (из самого «глубокого» состояния s_n вновь в него). Среднее время пребывания в ветви глубины n при среднем выигрыше W есть значение T_1 решения следующего разностного уравнения ($x=1, \dots, n$):

$$T_x = (pp_- + qp_+) T_{x-1} + (pq_- + qq_+) T_{x+1} + 1$$

с граничным условием $T_0 = 0$. Упростим запись уравнения, обозначив

$$P = pp_- + qp_+, \quad Q = pq_- + qq_+.$$

Очевидно, выполнено равенство $P+Q=1$.

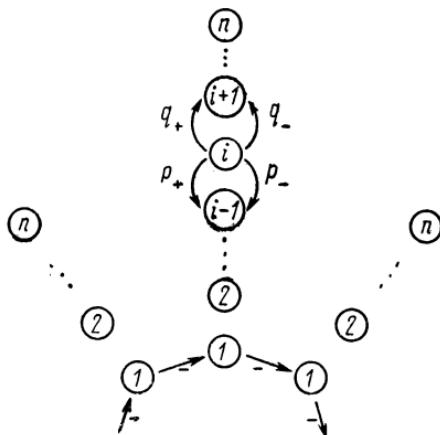


Рис. 4.

Уравнение (с граничным условием $T_0 = 0$)

$$T_x = PT_{x-1} + QT_{x+1} + 1 \quad (3)$$

сначала решаем в предположении $P \neq Q$. Общее решение имеет вид ($\lambda = P/Q$)

$$T_x = c_1 + c_2 \lambda^x + \frac{x}{P-Q},$$

c_1, c_2 — произвольные постоянные. Из граничного условия получаем, что $c_1 = -c_2$, а из уравнения (3) при $x=n$

$$T_n = T_{n-1} + \frac{1}{P}.$$

Подставляя сюда выражение для T_x , находим значение c_2 . Решение уравнения (3) оказывается следующим:

$$T_x = q \frac{1 - \lambda^x}{(P - Q)^2} \lambda^{-n} + \frac{x}{P - Q}. \quad (4)$$

Значит, интересующее нас среднее время равно

$$T^{(n)} = \frac{(Q/P)^n - 1}{Q - P}. \quad (5)$$

Отметим, что при $P = Q$ уравнение имеет вид

$$T_x = \frac{T_{x-1} + T_{x+1}}{2} + 1$$

и его решением служит $T_x = 2nx - x^2$. Отсюда среднее время равно

$$T^{(n)} = 2n - 1.$$

Исследованию квазилинейных автоматов мы предположим рассмотрение двух частных случаев.

Автоматы $K_{k,n}$ при поощрениях сдвигаются по ветви на один шаг «влубь» ($q_+ = 1$, $p_+ = 0$), а при наказании с равными вероятностями ($q_- = p_- = 1/2$) на один шаг влево или вправо. Средние времена пребывания в ветвях находятся из (5), где следует положить $P = p/2$ и $Q = q + p/2$. Тогда

$$T^{(n)}(W) = \frac{2}{1+W} \left[\left(\frac{3+W}{1-W} \right)^n - 1 \right].$$

Ясно (по теореме 3), что автоматы $K_{k,n}$ образуют ϵ -оптимальное семейство. Из рассмотрения выражения предельного среднего выигрыша видно, что с ростом n этот выигрыш стремится к максимуму экспоненциально быстро. Соответствующие рассуждения, аналогичные приведенным для автоматов $D_{k,n}$, мы воспроизвести не будем.

Воспользуемся решением уравнения (2) для построения еще одной модификации квазилинейных автоматов. В автоматах $K_{k,n}$ изменим функцию переходов, когда на входе — наказания.

Автоматы $L_{k,n}$ («с линейной тактикой») при поощрениях переходят от состояний с меньшими номерами (на данной ветви) к состояниям с номером, на единицу боль-

шим, а при наказаниях движутся в обратном направлении со сменой действия в состояниях s'_1 (циклически либо равновероятно). Очевидно, среднее время T_x удовлетворяет уравнению (3), в котором $P=p$, $Q=q$. Согласно (5) среднее время совершения действия автоматом $L_{k,n}$ равно

$$T^{(n)}(W) = \frac{\left(\frac{1+W}{1-W}\right)^n - 1}{W}.$$

Обратим внимание на то, что условие ϵ -оптимальности (теорема 3) не выполнено при произвольных W' и W'' , а лишь при $W'' \geq 0$ (в случае равенства нулю следует воспользоваться равенством (6)). Можно думать поэтому, что ϵ -оптимальность автоматов $L_{k,n}$ имеет место для тех бинарных ОПНЗ, которые удовлетворяют условию $\max_i W_i \geq 0$. Это предположение подтверждается прямым анализом предельного среднего выигрыша

$$W(L_{k,n}) = \frac{\sum_{i=1}^k \left[\left(\frac{1+W_i}{1-W_i} \right)^n - 1 \right]}{\sum_{i=1}^k \frac{\left(\frac{1+W_i}{1-W_i} \right)^n - 1}{W_i}}.$$

Стандартные вычисления дают следующий результат:

$$\lim_{n \rightarrow \infty} W(L_{k,n}) = \begin{cases} \max_i W_i, & \text{если } \max_i W_i \geq 0, \\ \frac{k}{\sum_1^k \frac{1}{W_i}}, & \text{если } \max_i W_i < 0. \end{cases}$$

Таким образом, в классе всех бинарных ОПНЗ автоматы $L_{k,n}$ не образуют ϵ -оптимального семейства. Это свойство справедливо в подклассе процессов, у которых $\max_i W_i \geq 0$, а для остальных процессов автоматы лишь «целесообразны», т. е. выполняется неравенство

$$W(L_{k,n}) \geq \frac{1}{k} \sum_{i=1}^k W_i, \quad W_i < 0, \quad i = 1, \dots, k,$$

причем знак равенства достигается на процессах, у которых все W_i равны.

Возвратимся к квазилинейным автоматам $Q_{k,n}$. Согласно формуле (5) их предельный средний выигрыш равен при условиях $P_i \neq Q_i$, $i = 1, \dots, k$,

$$W(Q_{k,n}) = \frac{\sum_{i=1}^k \frac{W_i \left(1 - \left(\frac{Q_i}{P_i}\right)^n\right)}{P_i - Q_i}}{\sum_{i=1}^k \frac{1 - \left(\frac{Q_i}{P_i}\right)^n}{P_i - Q_i}},$$

где использованы обозначения $P_i = p_- p_i + p_+ q_i$, $Q_i = q_- p_i + q_+ q_i$. Это выражение отличается от аналогичного $W(L_{k,n})$ предельного среднего выигрыша для автоматов $L_{k,n}$ смыслом величин P_i и Q_i . Воспользуемся уже известными для $L_{k,n}$ результатами.

При выполнении условия $Q_{\max} = \max_i Q_i \geq 1/2$ квазилинейные автоматы образуют ϵ -оптимальное семейство, а при обратном неравенстве $Q_{\max} < 1/2$ не образуют такого семейства. Простые выкладки показывают, что в случае, когда вероятности переходов связаны ограничением $q_- + p_+ < 1$, автоматы $Q_{k,n}$ целесообразны (в определенном ранее смысле). Если же $q_- + p_+ \geq 1$, то эти автоматы не обладают свойством целесообразности. Принимая условие $q_- + p_+ < 1$, находим, что ($p_{\min} = \min_i p_i$)

$$Q_{\max} = q_- p_{\min} + (1 - p_+) (1 - p_{\min}).$$

Отсюда вытекает, что высказанное условие ϵ -оптимальности эквивалентно следующему:

$$p_{\min} \leq \frac{\frac{1}{2} - p_+}{1 - q_- - p_+}.$$

Последнее всегда выполняется при $q_- \geq 1$, ибо тогда правая часть больше единицы. Если $q_- < 1/2$, $p_+ < 1/2$,

то автоматы $Q_{k,n}$ образуют ϵ -оптимальное семейство лишь для класса ОПНЗ, подчиненного требованию

$$p_{\min} \leq \frac{\frac{1}{2} - p_+}{1 - q_- - p_+}.$$

Рассмотрим еще одно семейство конечных автоматов, не являющееся ϵ -оптимальным в классе всех бинарных ОПНЗ.

Автоматы $G_{k,n}$ («с гистерезисной тактикой») имеют такие же, как $L_{k,n}$, переходы на ветвях, отличаясь правилами смены действий. Выходными состояниями, как и ранее, служат состояния s_1^j , а входными s_n^j , $j=1, \dots, k$. При двух действиях $k=2$ граф автомата напоминает петлю гистерезиса (отсюда название), при большем числе действий и циклической смене действий граф образует замкнутую линию. Вычисление среднего времени совершения действия снова опирается на уравнение (3) и является значением его решения (4) при $x=n$. Это означает, что

$$T^n(W) = \frac{1+W}{2W^2} \left[\left(\frac{1+W}{1-W} \right)^n - 1 \right] - \frac{n}{W},$$

если $W \neq 0$, и $T^{(n)} = n^2$, если $W=0$. Из анализа предельного среднего выигрыша

$$W(G_{k,n}) = \frac{\sum_{i=1}^k \left\{ \frac{1+W_i}{2W_i} \left[\left(\frac{1+W_i}{1-W_i} \right)^n - 1 \right] - n \right\}}{\sum_{i=1}^k \left\{ \frac{1+W_i}{2W_i^2} \left[\left(\frac{1+W_i}{1-W_i} \right)^n - 1 \right] - \frac{n}{W_i} \right\}}$$

при растущем числе состояний следуют выводы, аналогичные сделанным для автоматов $L_{k,n}$.

Мы рассмотрели несколько разных конструкций адаптивных систем. Зададимся вопросом о лучшей среди них. Для ответа на него следует выбрать критерий оптимальности. Естественно в качестве его принять скорость приближения предельного среднего выигрыша $W(A_n)$ к мак-

симальному значению. Ограничиваюсь автоматами с двумя действиями, определим величину

$$\Delta = \frac{W(A_n) - W_2}{W_1 - W_2},$$

где мы приняли $W_1 > W_2$ (т. е. y_1 — оптимальное действие). Мы знаем, что во всех изученных случаях $\Delta \rightarrow 1$ при $n \rightarrow \infty$ (для автоматов L и G при дополнительном условии $q_1 > 1/2$).

Сопоставим величины $\Delta(D)$ и $\Delta(L)$ столь разных ϵ -оптимальных семейств как $D_{2,n}$ и $L_{2,n}$, из которых $L_{2,n}$ не для произвольных ОПНЗ достигают цели (ϵ -оптимизации среднего выигрыша). Поэтому можно думать, что $D_{2,n}$ «лучше», чем $L_{2,n}$ во всех случаях, иными словами, априори кажется правдоподобной справедливость неравенства $\Delta(D) > \Delta(L)$. Проверим это, считая $q_1 > 1/2$.

Имеем

$$\Delta(D) = \frac{1}{1 + \left(\frac{p_1}{p_2}\right)^n}, \quad \Delta(L) = \frac{1}{1 + \gamma \left(\frac{p_1}{p_2}\right)^n},$$

где

$$\gamma = \left(\frac{q_2}{q_1}\right)^n \frac{W_1}{W_2} \frac{1 - \left(\frac{p_2}{q_2}\right)^n}{1 - \left(\frac{p_1}{q_1}\right)^n}.$$

Перечислим возможные типы соотношений $\Delta(D)$ и $\Delta(L)$ в зависимости от величины «суммарной» вероятности выигрыша $q_1 + q_2$.

1. Сначала примем $q_1 + q_2 = 1$. Легко убедиться, что тогда $\gamma = 1$ и, следовательно,

$$\Delta(D) = \Delta(L).$$

Автоматы $D_{2,n}$ и $L_{2,n}$ оказываются одинаково «хорошими» (подчеркнем, что сравниваются автоматы с одинаковой глубиной памяти n).

2. Далее, положим $q_1 + q_2 < 1$. Тогда имеем $q_2 < 1/2$ и $p_2 > q_1$ (т. е. $W_2 < 0$), поэтому

$$\gamma \sim \left(\frac{p_2}{q_1}\right)^n \frac{W_1}{-W_2} > 1.$$

Значит, в рассматриваемом случае

$$\Delta(D) > \Delta(L).$$

3. Наконец, пусть $q_1 + q_2 > 1$. Возможны такие два варианта (имеются в виду большие значения n):

а) $q_2 < 1/2$. В силу неравенства $p_2 < q_1$

$$\gamma \sim \left(\frac{p_2}{q_1}\right)^n \frac{W_1}{-W_2} < 1;$$

б) $q_2 \geqslant 1/2$. В силу $q_1 > q_2$

$$\gamma \sim \left(\frac{q_2}{q_1}\right)^n \frac{W_1}{W_2} < 1.$$

В обоих вариантах оказывается $\Delta(D) < \Delta(L)$.

Итак, сравнение автоматов $D_{2,n}$ и $L_{2,n}$ показывает, что ни один из них не может считаться лучшим другого, а именно, для процессов одного типа преимущество за L , другого типа — за D и третьего — они оказываются равнозначными.

Эта ситуация — типичная для адаптивных систем, и в дальнейшем мы не станем выяснять, какие адаптивные системы «оптимальные».

Число различных ϵ -оптимальных семейств конечных автоматов можно увеличивать беспрепятственно. Известные нам конструкции автоматов позволяют строить новые семейства, например семейство автоматов $(KD)_{k,n}$, у которых переходы при поощрениях таковы же, как у автоматов $D_{k,n}$, а при наказаниях — как у $K_{k,n}$.

§ 3. Конечные автоматы для ОПНЗ.

В этом параграфе мы рассмотрим еще несколько типов ϵ -оптимальных семейств автоматов. Они основаны на том же, что и в § 2, принципе преимущественного совершенствования тех действий, за которые чаще поступают поощрения. Отличие их от рассмотренных в предыдущем параграфе автоматов заключается в отказе от одновходовости и одновыходовости, от линейности графа. Мы ограничимся здесь лишь указанием конструкций автоматов, не вдаваясь в доказательства, которые стандартно проводятся

решением систем линейных уравнений для предельных вероятностей состояний

$$\pi^{(n)} \mathcal{P}^{(n)} = \pi^{(n)}$$

с обычным условием нормировки компонент вектора $\pi^{(n)}$, а затем проверкой того, что с ростом памяти n предельный средний выигрыш стремится к максимуму \bar{W} .

Рассмотрим автоматы «с сравнивающей цепочкой», являющиеся многовходовыми и многовыходовыми. Схематически они изображены в случае $k=2$ на рис. 5. Они состоят из трех блоков, два из которых, отмеченные символами Y_1 и Y_2 , представляют собой линейные ветви состояний того же типа, что и в предыдущем параграфе, а средний (отмеченный парой (y_1, y_2)) называется *сравнивающей цепочкой*. Последняя сопоставляет «полезность» обоих действий и «отбирает» лучшее из них (за которое

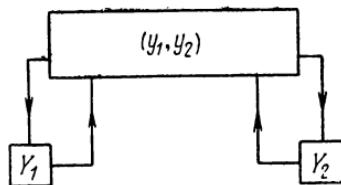


Рис. 5.

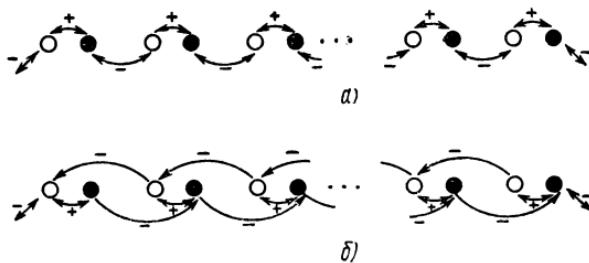


Рис. 6.

чаще приходят поощрения), т. е. если таковым оказалось y_i , осуществляется переход в блок Y_i .

Приведем два варианта сравнивающих цепочек. Граф первой из них V_m показан на рис. 6, а. Светлыми кружками обозначены состояния, которым сопоставлено y_1 , а темными y_2 . В блоке содержится $2m$ состояний, из которых s_1 и s_{2m} входные и выходные для блока: через первое осуществляется обмен с блоком Y_1 , а через второе — с Y_2 .

Мы видим, что если действие $y_1(+)$ влечет более частое, чем $y_2(-)$, появление поощрений, то при всяком начальном состоянии в цепочки (y_1, y_2) возникает случайное блуждание с преимущественным смещением влево к состояниям с меньшими номерами. В крайнем левом состоянии s_1 совершается действие y_2 , за которое с большой вероятностью появляется наказание, и автомат переходит в состояния блока Y_1 .

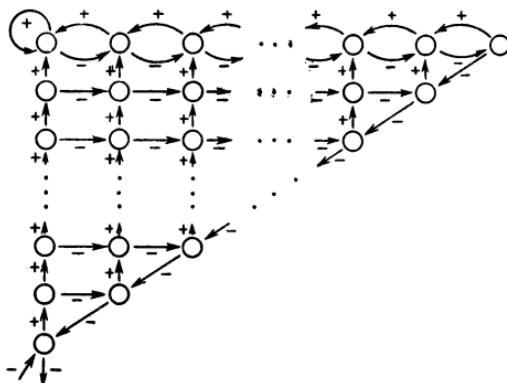


Рис. 7.

Примем за ветви Y_1 и Y_2 такие же ветви, как у автомата с гистерезисной тактикой. Нетрудно вычислить предельный средний выигрыш, вид которого приводит к заключению, что эти автоматы образуют ϵ -оптимальное семейство в классе всех бинарных ОПНЗ при неограниченно растущих n и m .

Граф другого варианта сравнивающей цепочки показан на рис. 6, б. Поощрения здесь приводят к смене действия, а серия наказаний вызывает движение по состояниям, отвечающим неблагоприятному действию, вплоть до перехода в блок, в котором избирается другое действие. Можно убедиться, что автоматы с таким строением сравнивающей цепочки образуют по индексам n и m ϵ -оптимальное семейство в классе всех бинарных ОПНЗ.

Нетрудно построить автоматы со сравнивающими действиями в случае более чем с двумя действиями.

Не обязательно, чтобы ϵ -оптимальные семейства автоматов описывались набором линейных графов. На рис. 7

изображен плоский граф подавтомата, соответствующего одному действию. Символ «+» означает поощрение x_1 , а «—» наказание x_2 . Автоматы с такого вида подавтоматами достигают цели благодаря «инерционности»: попав в самое глубокое состояние (по крайней мере за n шагов подряд), автомат покидает эти состояния не менее чем за $2n$ шагов.

§ 4. Автоматные алгоритмы управления ОПНЗ (не бинарный случай)

Обозначим через $K_{a, b; k}$ класс ОПНЗ со значениями в $X=[a, b]$ — конечном отрезке на числовой прямой и с пространством управлений $Y=\{y_1, \dots, y_k\}$. Назначим целью управления ϵ -оптимальность в сильном смысле. Для построения соответствующих адаптивных систем мы хотим воспользоваться изученными выше очень простыми конечными автоматами. Это на самом деле возможно сделать, если суметь трансформировать значения процесса в элементы входного бинарного алфавита автоматов.

Мы добавляем к автомatu с бинарным входом $A_{k, n}$ еще один блок — входной преобразователь, который действует следующим образом: получив значение процесса $\xi \in [a, b]$, он с вероятностью $\frac{\xi - a}{b - a}$ вырабатывает сигнал x_1 (поощрение) и с вероятностью $\frac{b - \xi}{b - a}$ — сигнал x_2 (наказание). Любой из них поступает в $A_{k, n}$ и вызывает надлежащую смену состояний, а с нею и очередной выходной сигнал — управление $y \in Y$.

Условное распределение вероятностей ОПНЗ ξ , обозначим $F(z|y)=\text{Вер}\{\xi < z|y\}$. Вычислим вероятность поощрения x_1 на входе автомата $A_{k, n}$: вероятность процессу ξ_t оказаться меньше z равна $F(z|y)$, а входной преобразователь образует x_1 с вероятностью $\frac{z - a}{b - a}$. Значит,

$$p(x_1|y) = \int_a^b \frac{z - a}{b - a} F(dz|y) = \frac{W(y) - a}{b - a}.$$

Аналогично находим $p(x_2|y)$.

В качестве $A_{k,n}$ могут фигурировать любые автоматы из числа ранее рассмотренных. Условимся обозначать полученные из них адаптивные системы для не бинарных ОПНЗ тем же символом, что был употреблен в §§ 2, 3, но с тильдой наверху. Так, избрав автомат $D_{k,n}$, соответствующую систему обозначим $\tilde{D}_{k,n}$. Мы не будем в обозначении указывать допустимую область процесса — отрезок $[a, b]$, считая это ясным каждый раз из контекста.

Системы $\tilde{A}_{k,n}$, $n \geq 1$, образуют ε -оптимальное в сильном смысле семейство для описанных классов ОПНЗ.

Рассмотрим еще один подход к автоматам, управляющим не бинарными ОПНЗ.

Через $O_{k,\alpha}$ обозначим вероятностные автоматы Мура, изображаемые набором

$$O_{k,\alpha} = \{X, S, Y; \Pi_x, q\},$$

где $X = \{x_1, \dots, x_m\}$ — конечное числовое множество входных сигналов, $Y = \{y_1, \dots, y_k\}$ — множество действий, $S = \{s_1, \dots, s_k\}$ — множество состояний. Функция выходов q — детерминированная: в состоянии s_i автомат совершает действие y_i . Функция переходов задана стochастическими матрицами

$$\Pi_x = (p_{ij}(x)),$$

элементы которых определены равенствами

$$p_{ii}(x) = 1 - g_\alpha(x),$$

$$p_{ij}(x) = \frac{g_\alpha(x')}{k-1}, \quad i \neq j,$$

где $g_\alpha(x)$ — функция на множестве X со значениями на $(0, 1)$, удовлетворяющая при любых $x', x'', x' > x''$, неравенству

$$\frac{g_\alpha(x')}{g_\alpha(x'')} < \alpha, \quad \alpha \in (0, 1).$$

Назовем такие автоматы *оценывающими*.

Зададимся числовой последовательностью $\alpha_n \rightarrow 0$ и рассмотрим отвечающую ей последовательность автоматов O_{k,α_n} .

Сначала допустим, что эти автоматы управляют ОПНЗ с вырожденными распределениями вероятностей, т. е. в ответ на действие y_i процесс принимает значение x_i , $i=1, \dots, k$. Покажем, что при $\alpha_n \rightarrow 0$ предельная вероятность действия, за которое причитается максимальный выигрыш (наибольший элемент из X), сходится к 1.

Состояния автомата O_{k, α_n} образуют марковскую цепь, у которой переходные вероятности p_{ij} при $i \neq j$ все одинаковы. Нетрудно убедиться, что предельные вероятности состояний такой цепи существуют и удовлетворяют равенству

$$\frac{\pi_i}{\pi_j} = \frac{p_{ji}}{p_{ij}}, \quad i \neq j.$$

Применим к O_{k, α_n} имеем $p_{ij} = \frac{1}{k-1} g_{\alpha_n}(x_i)$ и $p_{ji} = \frac{1}{k-1} g_{\alpha_n}(x_j)$. Из определения автоматов получаем высказанное утверждение.

Теперь предположим, что распределения управляемого ОПНЗ не вырождены. Тогда переходные вероятности введенной в предыдущем абзаце марковской цепи удовлетворяют неравенствам

$$c_1 g_{\alpha_n}(x_i) \leq p_{ij} \leq c_2 g_{\alpha_n}(x_i), \quad c_1, c_2 > 0,$$

где $x_i = \min_{l: p(x_l | y_i) > 0} x_l$ — минимальный выигрыш, приходящий с ненулевой вероятностью за действие y_i . Отсюда следует, что с вероятностью, стремящейся к 1 (при $\alpha_n \rightarrow 0$), избирается действие, которое максимизирует величину x_i .

Таким образом, оценивающие автоматы являются ϵ -оптимальным семейством по отношению к специфической — максминной — цели управления ОПНЗ.

§ 5. δ-оптимальные автоматы

Рассматривается класс $Q_{X, k}$ ОПНЗ ξ , с одними и теми же пространствами: фазовым пространством X — числовым и пространством управлений $Y = \{y_1, \dots, y_k\}$ — конечным. Изберем целью управления ϵ -оптимальность в силь-

ном смысле. Средством достижения этой цели является надлежащим образом организованное блуждание по множеству распределений на Y . Этим множеством является открытый $(k-1)$ -мерный симплекс

$$F = \{(p_1, \dots, p_k) : \sum p_i = 1, p_j > 0, j = 1, \dots, k\}$$

Смысл компоненты p_i вектора $\mathbf{p} = (p_1, \dots, p_k)$ — вероятность совершения действия y_i .

Пусть L — обучаемая система (с X и Y в качестве множеств входных и выходных сигналов), у которой множеством правил выбора действий является симплекс F . Зададим число $\delta \in (0, 1)$ и скажем, что в момент t система L находится в δ -оптимальном режиме, если распределение $\mathbf{p}(t)$ таково, что вероятность выбора оптимального действия *) равна $1 - \delta$. Сумма вероятностей всех остальных действий равна δ .

Обучаемая система L_δ называется δ -оптимальной, если для любого управляемого ею ОПНЗ она спустя время $\tau(\omega)$ (такое, что $P(\tau < \infty) = 1$) попадает в δ -оптимальный режим (и впредь не покинет) с фиксированным значением числа δ .

Если оптимальное действие единственно и имеет индекс j_0 , то при $t > \tau(\omega)$ вектор \mathbf{p} имеет вид

$$\mathbf{p}_\delta = (\delta q_1, \dots, \delta q_{j_0-1}, 1 - \delta, \delta q_{j_0+1}, \dots, \delta q_k) = (p_1^{(\delta)}, \dots, p_k^{(\delta)}),$$

где $q_j > 0$, $\sum_{j \neq j_0} q_j = 1$ и средний выигрыш за один такт равен

$$(1 - \delta) \bar{W} + \delta \sum_{j \neq j_0} q_j W(y_j).$$

Часто удобно вектор \mathbf{p}_δ задавать в виде $\mathbf{p}_\delta = \left(\frac{\delta}{k-1}, \dots, \frac{\delta}{k-1}, 1 - \delta, \frac{\delta}{k-1}, \dots, \frac{\delta}{k-1} \right)$, поэтому выражение

*) Напомним, что оптимальное действие y_0 приводит к максимуму среднего выигрыша $W(y_0) = \max_i W(y_i)$.

среднего выигрыша упрощается:

$$W_\delta(L_\delta) = \bar{W} - \delta \left(\bar{W} - \frac{1}{k-1} \sum_{j \neq j_0} W(y_j) \right).$$

Обратим внимание, что при малых δ приходим к свойству ϵ -оптимальности в слабом смысле. Это означает, что совокупность обучаемых систем L_δ образует ϵ -оптимальное (в слабом смысле) семейство. Мы установим более сильный результат.

Теорема 1. Совокупность δ -оптимальных обучаемых систем L_δ образует ϵ -оптимальное в сильном смысле семейство.

Доказательство. Рассмотрим средний выигрыш L_δ за время t :

$$V(t) = \frac{1}{t} \sum_{j=1}^t \xi_j = \sum_{i=1}^k \frac{N_i(t)}{t} V_i(t),$$

где $N_i(t)$ — количество совершений действия y_i , $V_i(t) = \frac{1}{N_i(t)} \sum_{l=1}^{N_i(t)} x_l^{(i)}$ — эмпирический средний выигрыш за действие y_i в течение отрезка времени $[1, t]$. При всех $t > \tau_\delta$ (ω) система L_δ оказывается в δ -оптимальном режиме. Значит, при $t \rightarrow \infty$ с вероятностью 1 имеем ($p_i^{(\delta)}$ — i -я компонента вектора \mathbf{p}_δ)

$$\frac{N_i(t)}{t} \rightarrow p_i^{(\delta)}, \quad V_i(t) \rightarrow W(y_i), \quad i = 1, \dots, k.$$

Отсюда заключаем, что для всех достаточно больших t с вероятностью 1

$$V(t) = \sum_1^k \frac{N_i(t)}{t} V_i(t) > W_\delta(L_\delta; \xi) - \frac{\epsilon}{2},$$

т. е. при малых δ , когда $W_\delta(L; \xi) > \bar{W} - \epsilon/2$, получаем $V(t) > \bar{W}_\delta - \epsilon/2 - \epsilon/2 = \bar{W} - \epsilon$.

Обратимся к задаче синтеза δ -оптимальных адаптивных систем. При конечном X и извлечении из F не более чем счетного подмножества элементов такие системы

представляют собой автоматы. Каковы возможности таких автоматов? На этот вопрос отвечает следующая теорема.

Теорема 2. *Не существует конечного δ -оптимального автомата для всех ОПНЗ с одинаковыми (конечными) X и Y .*

Доказательство. Пусть A — конечный вероятностный автомат Мура $A = (X, S, Y; s_0; \Pi_x, q)$, где $q = q(y | s)$ — условное распределение вероятностей на Y , т. е. семейство из не более чем $|S|$ различных распределений. Объекту $A \otimes \xi$ поставлена в соответствие ассоциированная марковская цепь (S, \mathcal{P}) . В ней распределениям p (на множестве Y) соответствуют некоторые множества состояний G_p . Обозначим через $G_{i, \delta}$ множества состояний, которым отвечают векторы вида

$$\begin{aligned} p_\delta = & (\delta q_1, \dots, \delta q_{i-1}, 1 - \delta, \delta q_{i+1}, \dots, \delta q_k), \\ \sum_{l \neq i} q_l = & 1, \quad q_l > 0. \end{aligned}$$

Если A δ -оптимальен (по отношению ко всем ОПНЗ с данными X и Y), то 1) все множества состояний $G_{i, \delta}$ непусты; 2) все множества $G_{i, \delta}$ цепи (S, \mathcal{P}) сообщающиеся; 3) для ОПНЗ, у которого действие y_{i_0} оптимально (i_0 произвольно), группа состояний $G_{i_0, \delta}$ поглощающая. Последнее означает, что спустя конечное время автомат навсегда останется в состояниях из $G_{i, \delta}$. Условия 2) и 3) несовместимы с требованием конечности марковской цепи. Теорема доказана.

Доказательство существования δ -оптимальных автоматов производится их построением. Мы сначала приведем, однако, пример «почти» δ -оптимального автомата. В его определении участвуют k -мерные векторы N и V , уже введенные нами выше,

$$N = (N_1, \dots, N_k), \quad V = (V_1, \dots, V_k),$$

где $N_i = N_i(t)$ — количество совершений действия y_i (за время t), а $V_i = V_i(t)$ — эмпирический средний выигрыш за действие y_i (в течение времени t). Автомат M_δ на каждом такте пересчитывает \bar{N} и \bar{V} и выявляет «лучшее» действие (на него приходится наибольший эмпирический средний выигрыш). Выберем число $r \geq 1$ и подпоследова-

тельность натурального ряда (n_i) , $n_i < n_{i+1}$. Способ выбора действий заключается в чередовании двух режимов: а) по r раз подряд выполняется каждое из k действий, б) на протяжении n_i тактов лучшее действие избирается с вероятностью $1 - \delta$ (остальные избираются равновероятно с суммарной вероятностью δ). Спустя достаточно большое время, когда эмпирический средний выигрыш V_{j_0} за оптимальное действие y_{j_0} превысит остальные величины V_j , $j \neq j_0$, автомат M_δ подавляющую долю времени, стремящуюся со временем к 1, будет проводить в δ -оптимальном режиме.

Обратимся к заданию класса автоматов, называемых δ -автоматами, относительно которых будет доказана δ -оптимальность.

Рассматриваются автоматы

$$A = (X \times Y, S, Y; s_0; \sigma, \eta).$$

У них входными сигналами служат пары (x, y) , во времени это (ξ_t, y_{t-1}) , состояниями — тройки $s = (N, V, p)$, начальное состояние $s_0 = (N=V=0, p=(1/k, \dots, 1/k))$. Изменение компонент состояния N и V происходит очевидным образом (в момент t пересчитываются i -е компоненты этих векторов, где i — номер действия y_{t-1}). Условимся считать при $N_i(t)=0$, что $V_i(s)=0$ при всех $s \leq t$.

Функция переходов σ автомата представляет собой пару преобразований $\sigma = (\mathcal{O}J)$. Здесь \mathcal{O} — оператор, который воздействует на распределение p в моменты времени, указываемые правилом J . Мы допустим, что это правило ограничено требованиями

$$\mathsf{P}(\lim_{t \rightarrow \infty} v_t = \infty) = 1,$$

где v_t — количество преобразований вектора p за время t ; конечны все моменты случайной величины Δ — интервала времени между двумя последовательными приложениями оператора \mathcal{O} .

Вот несколько примеров правил J : 1) в каждый момент времени вектор p трансформируется с вероятностью $\beta > 0$, а с вероятностью $1 - \beta$ не меняется; 2) p преобразуется после того, как на 1 возрастет наименьшее из

чисел N_j ; 3) между двумя последовательными трансформациями \mathbf{p} совершаются все возможные действия.

Оператор O переводит стохастический вектор \mathbf{p} снова в стохастический $O\mathbf{p}$ и подчинен условиям:

1. Все компоненты $O\mathbf{p}$ ограничены снизу числом $\alpha > 0$.
2. Пусть j_0 — номер лучшего действия в момент очередного преобразования \mathbf{p} , т. е. $V_{j_0}(t) = \max V_j(t)$. Тогда вектор $O\mathbf{p}$ удовлетворяет неравенству $||O\mathbf{p} - e_{j_0}|| \leq ||\mathbf{p} - e_{j_0}||$, где $e_{j_0} = (0, \dots, 0, 1, 0, \dots, 0)$ — j_0 -вершина симплекса F , а $||\cdot||$ — евклидова норма.

3. При сохранении y_{j_0} в качестве лучшего действия в течение не более T_δ воздействий оператора O результатом итераций O становится δ -оптимальный режим, при котором вероятность 1— δ приходится на j_0 -ю компоненту. Согласно условию на правило выбора моментов приложения оператора O отсюда следует, что описанная процедура занимает время, не превосходящее $\Delta_1 + \dots + \Delta_{T_\delta}$. Все моменты этой случайной величины конечны.

4. При смене «лучшего» действия y_{j_0} на y_{i_0} в момент очередной трансформации \mathbf{p} преобразованию подвергается вектор \mathbf{p}_{i_0} , который получается из имеющегося вектора \mathbf{p}_{j_0} перестановкой его j_0 -й и i_0 -й компонент.

Заметим, что при управлении автоматом A процессами с единственным оптимальным действием условие 4 можно ослабить. Например, при смене лучшего действия допустимо возвратиться к исходному, равномерному на Y распределению.

Из высказанных допущений относительно трансформаций \mathbf{p} следует, что вероятность лучшего действия монотонно растет, приближаясь к максимальному значению 1— δ . Это очевидно из условий 2 и 3 при сохранении «лучшего» действия, а при смене его следует из того, что в результате перестановки компонент новому «лучшему» действию присваивается уже накопленная большая вероятность. Поэтому за конечное число итераций оператора O одна из компонент \mathbf{p} станет равной 1— δ .

Функцией выходов автомата A служит распределение \mathbf{p} на множестве действий Y . Иными словами, A является вероятностным автоматом Мура.

Трудности аналитического исследования объекта $A \otimes \xi$ выясняет следующая теорема.

Теорема 3. При конечном фазовом пространстве X ОПНЗ ξ_t , управляемого $\delta\omega$ -автоматом A , с не более чем счетным множеством распределений p , ассоциированная марковская цепь (S, \mathcal{P}) счетная и все ее состояния невозвратные.

Доказательство. Состояниями цепи (S, \mathcal{P}) служат тройки (N, V, p) , где векторы V и p принимают не свыше счетного множества значений, а N в точности счетно. Значит, эта цепь счетная.

Пусть в момент t_1 цепь оказалась в состоянии $s(t_1) = (N(t_1), V(t_1), p(t_1))$. При $t_2 > t_1$ вектор $N(t_2)$ отличается от $N(t_1)$ по крайней мере одной компонентой. Из компонент $N(t_2) - N(t_1)$ образуем сумму, она равна $t_2 - t_1 > 0$. Значит, ни при каком t текущее состояние s_t не может совпасть ни с одним прошлым состоянием. Теорема доказана.

Исследование $\delta\omega$ -автоматов, как адаптивных систем для ОПНЗ, базируется на установлении свойств компонент марковской цепи (S, \mathcal{P}) — процессов $N(t)$, $V(t)$, $p(t)$. Начнем с процесса $N(t)$.

Значения процесса $N(t)$ принадлежат целочисленной k -мерной решетке. В силу тождества $N_1(t) + \dots + N_k(t) = t$, его компоненты линейно зависимы, в случайному порядке возрастают на единицу, причем вероятность p_j возрастания $N_j(t)$ лежит в пределах $0 < \alpha \leq p_j \leq 1 - \delta$. Легко видеть, что компоненты образуют полумартингал. Нас интересуют оценки вероятности больших уклонений компонент $N_i(t)$ от математических ожиданий. Они опираются на следующий результат.

Лемма 1. Пусть η_1, \dots, η_n — независимые случайные величины, $a_j \leq \eta_j \leq b_j$, $j=1, \dots, n$. Тогда для любого $x > 0$

$$\Pr \left(\sum_{i=1}^n \eta_i - \sum_{i=1}^n E\eta_i \geq nx \right) \leq \exp \left(- \frac{2x^2}{\sum_{j=1}^n (b_j - a_j)^2} n^2 \right).$$

Теперь получим оценки для $N_i(t)$.

Лемма 2. При всех t , любом $\varepsilon > 0$ и $j = \overline{1, k}$ верны неравенства

$$\mathbf{P}(N_j(t) \leqslant (\alpha - \varepsilon)t) \leqslant e^{-2\varepsilon^2 t},$$

$$\mathbf{P}(N_j(t) \geqslant (1 - \delta + \varepsilon)t) \leqslant e^{-2\varepsilon^2 t}.$$

Доказательство проведем для второго неравенства. Обозначим через $N^*(t)$ случайный процесс вида $N^*(t) = \zeta_1 + \dots + \zeta_t$, где

$$\zeta_i = \begin{cases} 1 & \text{с вероятностью } 1 - \delta, \\ 0 & \text{с вероятностью } \delta. \end{cases}$$

Очевидно, $\mathbf{P}(N^*(t) \geqslant Lt) \geqslant \mathbf{P}(N_j(t) \geqslant Lt)$. Согласно лемме 1, примененной к величинам (ζ_i) , которые удовлетворяют неравенствам $0 \leqslant \zeta_i \leqslant 1$ и имеют математическое ожидание $E\zeta_i = 1 - \delta$,

$$\begin{aligned} \mathbf{P}(N^*(t) \geqslant (1 - \delta + \varepsilon)t) &= \mathbf{P}(N^*(t) - (1 - \delta)t \geqslant \varepsilon t) = \\ &= \mathbf{P}(N^*(t) - EN^*(t) \geqslant \varepsilon t) \leqslant e^{-2\varepsilon^2 t}. \end{aligned}$$

Отсюда вытекает одно утверждение леммы, другое устанавливается аналогично.

Лемма 3.

$$\mathbf{P}\left(\alpha < \liminf_{t \rightarrow \infty} \frac{N_j(t)}{t}, \limsup_{t \rightarrow \infty} \frac{N_j(t)}{t} \leqslant 1 - \delta, j = \overline{1, \dots, k}\right) = 1.$$

Доказательство. При любом $\varepsilon > 0$ ряд $\sum \mathbf{P}(N_j(t) \leqslant (\alpha - \varepsilon)t)$ сходится и по лемме Бореля—Кантелли с вероятностью 1 события $\left\{\frac{N_j(t)}{t} \leqslant \alpha - \varepsilon\right\}$ наступают лишь конечное число раз. То же самое верно для событий $\left\{\frac{N_j(t)}{t} \geqslant 1 - \delta + \varepsilon\right\}$. Это доказывает лемму.

Перейдем к исследованию свойств компонент $V_j(t)$ вектора эмпирических средних выигрышей. Следует преодолеть затруднение, состоящее в том, что $V_j(t)$ являются средними арифметическими случайного числа случайных слагаемых. Прежде всего заметим, что из соотношения $N_j(t) \sim \lambda t$ ($\alpha \leqslant \lambda \leqslant 1 - \delta$) и состоятельности среднего

арифметического как оценки математического ожидания вытекает усиленный закон больших чисел.

Л е м м а 4.

$$\mathbb{P}\left(\lim_{t \rightarrow \infty} V_j(t) = W(y_j), \ j = \overline{1, k}\right) = 1,$$

Все k последовательностей $V_j(t)$ асимптотически нормальны с параметрами $(W(y_j), \sigma(y_j)/\sqrt{N_j(t)})$, причем $N_j(t) \sim \lambda t$.

Нам потребуются вспомогательные неравенства. Сначала их сформулируем применительно к управляемым ОПНЗ с единственным оптимальным действием.

Л е м м а 5. Пусть среди k имеющихся действий y_1 — единственное оптимальное. Тогда

$$\mathbb{P}(V_1(t) \leqslant \max_{i=2, \dots, k} V_i(t)) < (k-1)(t^2 + 8)e^{-\mu t}, \quad \mu > 0.$$

Д о к а з а т е л ь с т в о. Согласно формуле полной вероятности

$$\begin{aligned} \mathbb{P}(V_1(t) \leqslant \max_{i>1} V_i(t)) &= \sum_{j=2}^k \sum_{l_i, l_j=0}^t \mathbb{P}(V_1(t) \leqslant V_j(t) \mid V_j(t) = \\ &= \max_{i>1} V_i(t); \quad N_1(t) = l_1, \quad N_j(t) = l_j) \mathbb{P}(V_j(t) = \\ &= \max_{i>1} V_i(t) \mid N_j(t) = l_j) \mathbb{P}(N_1(t) = l_1, \quad N_j(t) = l_j). \end{aligned}$$

Заменяя единицей второй множитель в сумме, получаем

$$\begin{aligned} \mathbb{P}(V_1(t) \leqslant \max_{i>1} V_i(t)) &< \sum_{j=1}^k \sum_{l_i, l_j=0}^t \mathbb{P}(V_1(t) \leqslant V_j(t) \mid V_j(t) = \\ &= \max_{i>1} V_i(t); \quad N_1(t) = l_1, \quad N_j(t) = l_j) \times \\ &\quad \times \mathbb{P}(N_1(t) = l_1, \quad N_j(t) = l_j). \quad (1) \end{aligned}$$

Во внешней сумме по индексу j выберем одно слагаемое, например, соответствующее $j=2$. Действиям y_1 и y_2 отвечают случайные $\xi(y_1)$ и $\xi(y_2)$, задаваемые распределе-

ниями $\mu(M | y_i)$, с математическими ожиданиями W_1 и W_2 . Представим это слагаемое в виде

$$\sum_{l_1, l_2=0}^t \mathbb{P}(V_1(t) - V_2(t) \leq 0 | N_i(t) = l_i, i=1, 2) \times \\ \times \mathbb{P}(N_i(t) = l_i, i=1, 2) = \Sigma_1 + \Sigma_2.$$

Суммирование в Σ_1 ведется по значениям $N_i(t)$, подчиненным неравенствам $(\alpha - \varepsilon)t \leq N_i(t) \leq (1 - \delta + \varepsilon)t$, $i=1, 2$, а в Σ_2 — по остальным значениям индексов. Оценим Σ_1 сверху.

Сначала заметим, что

$$\Sigma_1 < \sum_{l_1, l_2=(\alpha-\varepsilon)t}^{(1-\delta+\varepsilon)t} \mathbb{P}\left(\frac{1}{l_1} \sum_{i=1}^{l_1} x_i^{(1)} - \frac{1}{l_2} \sum_{i=1}^{l_2} x_i^{(2)} \leq 0\right),$$

где (x_i^*) — значения случайной величины $\xi(y_x)$, $x=1, 2$. Рассмотрим одно из слагаемых справа. Полагая $l_i = \lambda_i t$ ($\alpha - \varepsilon \leq \lambda_i \leq 1 - \delta + \varepsilon$) и вводя центрированные случайные величины $z^{(*)} = x^{(*)} - W_x$, сделаем еще одну замену

$$\zeta_j = \frac{z_j^{(1)}}{\lambda_1}, \quad j=1, \dots, \lambda_1 t; \quad \zeta_{\lambda_1 t+j} = -\frac{z_j^{(2)}}{\lambda_2}, \quad j=1, \dots, \lambda_2 t.$$

Очевидно, что $\zeta_j \in \left[\frac{a}{\lambda_1}, \frac{b}{\lambda_1}\right]$, $j=1, \dots, \lambda_1 t$; $\zeta_{\lambda_1 t+j} \in \left[-\frac{b}{\lambda_2}, -\frac{a}{\lambda_2}\right]$, $j=1, \dots, \lambda_2 t$ и $E\zeta_j = 0$. Для суммы $S_t = \sum_{j=1}^{\lambda_1 t} \zeta_j$

имеем согласно лемме 1, принимая во внимание, что $ES_t = 0$,

$$M = W_1 - W_2 > 0,$$

$$\mathbb{P}\left(\frac{1}{l_1} \sum_{i=1}^{l_1} x_i^{(1)} - \frac{1}{l_2} \sum_{i=1}^{l_2} x_i^{(2)} \leq 0\right) = \mathbb{P}(S_t \leq -Mt) = \\ = \mathbb{P}\left(S_t \leq -\frac{M}{\lambda_1 + \lambda_2} (\lambda_1 + \lambda_2) t\right) \leq \exp\left\{-2 \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \frac{M^2}{(b-a)^2} t\right\}. \quad (2)$$

Первый множитель под знаком экспоненты зависит от выбора слагаемого в сумме \sum_1 . При наименьшем его значении получаем мажоранту каждой входящей в \sum_1 вероятности

$$\mathbb{P}(S_t \leq -Mt) \leq \exp \left\{ -\frac{(\alpha - \varepsilon)^2}{1 - \delta + \varepsilon} \frac{M_0^2}{(b - a)^2} t \right\} = e^{-\mu t},$$

где $M_0 = \min_{j>1} (W_1 - W_j)^2$. Теперь легко оценивается сумма \sum_1 :

$$\sum_1 < \sum_{l_1, l_2=(\alpha-\varepsilon)t}^{(1-\delta+\varepsilon)t} \mathbb{P} \left(\frac{1}{l_1} \sum_{i=1}^{l_1} x_i^{(1)} - \frac{1}{l_2} \sum_{i=1}^{l_2} x_i^{(2)} \leq 0 \right) < t^2 e^{-\mu t}.$$

Остается оценить сумму \sum_2 :

$$\begin{aligned} \sum_2 &< \mathbb{P}(N_1(t) \leq (\alpha - \varepsilon)t) + \mathbb{P}(N_2(t) \leq (\alpha - \varepsilon)t) + \\ &+ \mathbb{P}(N_i(t) \leq (\alpha - \varepsilon)t, i = 1, 2) + \mathbb{P}(N_1(t) \geq (1 - \delta + \varepsilon)t) + \\ &+ \mathbb{P}(N_2(t) \geq (1 - \delta + \varepsilon)t) + \\ &+ \mathbb{P}(N_i(t) \geq (1 - \delta + \varepsilon)t, i = 1, 2) \leq \\ &\leq 4\mathbb{P}(N(t) \leq (\alpha - \varepsilon)t) + 4\mathbb{P}(N(t) \geq (1 - \delta + \varepsilon)t). \end{aligned}$$

Согласно неравенствам леммы 2 приходим к следующему:

$$\sum_2 < 8e^{-2\varepsilon^2 t}.$$

Объединяя полученные результаты, находим

$$\sum_1 + \sum_2 < t^2 e^{-\mu_1 t} + 8e^{-2\varepsilon^2 t} < (t^2 + 8)e^{-\mu t}, \quad (3)$$

где $\mu = \min(\mu_1, 2\varepsilon^2)$. Доказательство леммы завершаем ссылкой на оценку

$$\mathbb{P}(V_1(t) \leq V_j(t)) < e^{-\mu t},$$

получаемую так же, как (2). Из нее непосредственно вытекает

$$\mathbb{P}(V_1(t) \leq \max_{j>1} V_j(t)) < (k-1)(t^2 + 8)e^{-\mu t}.$$

Лемма доказана.

Следствие 1. При достаточно больших t имеем

$$\mathbb{P} \left(V_1(t) \leqslant \max_{j>1} V_j(t) \right) < ae^{-\lambda t}, \quad a > 0, \lambda > 0.$$

Следствие 2. Если ОПНЗ имеет $l \geqslant 2$ оптимальных действий y_1, \dots, y_l , то

$$\mathbb{P} \left(\min_{i \leqslant l} V_i(t) \leqslant \max_{j > l} V_j(t) \right) < be^{-\lambda t}, \quad b > 0.$$

Полученные оценки опираются на лемму 1, не предполагающую конечность множества входных сигналов автомата. Поэтому сформулированные леммы и дальнейшие результаты справедливы для класса ОПНЗ, у которых фазовым пространством служит отрезок числовой прямой. Можно было бы воспользоваться какой-либо теоремой об экспоненциальной оценке больших уклонений, например теоремой С. Н. Бернштейна, относящейся к неограниченным величинам и требующей дополнительные условия на их распределения (точнее, на скорость роста моментов). Не вдаваясь сейчас в рассмотрение возможных вариантов таких теорем, заметим лишь, что оценки в лемме 5 (и, разумеется, в следствии 2) завышенные. Однако порядок (экспоненциальный) скорости убывания рассматриваемых вероятностей не улучшаем.

Лемма 6. При управлении $\delta\omega$ -автоматом ОПНЗ ξ_t , у которого действия y_1, \dots, y_l ($1 \leqslant l \leqslant k$) оптимальные, неравенство $\min_{i \leqslant l} V_i(t) > \max_{i > l} V_j(t)$ может нарушаться с вероятностью 1 только конечное число раз.

Доказательство вытекает из оценок вероятностей рассматриваемого события (следствие 2 из леммы 5) и леммы Бореля — Кантелли.

Введем теперь случайные величины: τ'_δ — момент первого попадания автомата в δ -оптимальный режим (это марковский момент), τ''_δ — момент последнего попадания автомата в этот режим (т. е. при $t > \tau''_\delta$ автомат никогда не покинет этот режим; это немарковский момент), а также в предположении, что y_1 — единственное оптимальное

действие,

$$\tau_1 = \min \left\{ t: V_1(t) > \max_{i>1} V_i(t) \right\},$$

$$\tau_n = 1 + \max \left\{ t: V_1(t) \leq \max_{i>1} V_i(t) \right\}.$$

Справедливы неравенства

$$\tau_1 \leq \tau'_\delta \leq \tau''_\delta \leq \tau_n + T_\delta \Delta, \quad (4)$$

где $T_\delta \Delta$ означает случайное время, за которые оператор O переведет распределение p в δ -окрестность вершины вероятностного симплекса, отвечающей оптимальному действию. Эти неравенства сохраняются при нескольких оптимальных действиях, если очевидным образом модифицировать определения величин τ_1 и τ_n .

Л е м м а 7. *Все моменты случайных величин τ_1 и τ_n конечны.*

Д о к а з а т е л ь с т в о. Из следствия 1 (леммы 5) находим оценку сверху вероятности $p_{\tau_1}(t) = P(\tau_1 = t)$. В самом деле,

$$p_{\tau_1}(t) = P(V_1(s) \leq \max_{i>1} V_i(s), s < t; V_1(t) > \max_{i>1} V_i(t)) \leq \\ \leq P(V_1(t-1) \leq \max_{i>1} V_i(t-1)) < ae^{-\lambda(t-1)}.$$

Следовательно, все моменты величины τ_1 конечны. Для τ_n проходит аналогичное рассуждение.

С помощью доказанных лемм установим свойства δ -автоматов.

Т е о р е м а 4. *δ-автоматы при управлении ОПНЗ обладают свойством δ-оптимальности.*

Д о к а з а т е л ь с т в о. Из сходимости (с вероятностью 1) оценок $V_j(t)$ к средним выигрышам $W(y_j)$ вытекает, что при всех достаточно больших t оценка $V_{y_{opt}}(t)$ становится наибольшей среди всех оценок. Тогда оператор O не более чем за T_δ итераций трансформирует распределение p к виду $(\dots, 1 - \delta, \dots)$, в котором $1 - \delta$ занимает компоненту с номером оптимального действия. В случае нескольких оптимальных действий условие 4 на оператор O гарантирует сохранение, начиная с некоторого момента, δ-оптимального режима.

Теорема 5. Все моменты случайных величин τ'_δ и τ''_δ конечны.

Доказательство непосредственно следует из леммы 7 и неравенств (4).

Теперь можно представить наглядно функционирование $\delta\omega$ -автомата A как адаптивной системы управления для ОПНЗ. В течение начального промежутка случайной длины τ'_δ автомат перебирает в случайном порядке действия и, накапливая эмпирические средние выигрыши, по максимальному из них отыскивает лучшее действие. В момент τ'_δ впервые оптимальному действию приписывается подавляющая вероятность $1 - \delta$, но затем автомат может покидать δ -оптимальный режим конечное число раз. Однако спустя время τ''_δ с начала управления он навсегда останется в этом режиме.

Пусть множество значений ОПНЗ принадлежит конечному отрезку. Тогда последовательность средних выигрышей $W(\delta)$ рассматриваемых автоматов равномерно по множеству процессов стремится (при $\delta \rightarrow 0$) к максимальному выигрышу.

Обратим внимание на то, что $\delta\omega$ -автоматы не решают задачи исследования динамических свойств объекта управления, не оценивают распределения вероятностей $\mu(M|y_i)$, $i=1, \dots, k$. Находятся лишь оценки средних выигрышей $W(y_i)$, с помощью которых порождается направленное блуждание на множестве p .

§ 6. Примеры $\delta\omega$ -автоматов

В предыдущем параграфе, характеризуя $\delta\omega$ -автоматы, мы оставили большой произвол в выборе оператора O , подчинив его лишь четырем условиям. Теперь введем подкласс $\delta\omega$ -автоматов, у которого оператор O определен с точностью до двух числовых параметров, имеющих ясный смысл.

Зададим числа θ , \hat{p} , δ , подчинив их условиям

$$0 < \theta < 1 - \frac{1}{k}, \quad \frac{\lfloor \theta k \rfloor}{k} < \hat{p} \leq 1 - \delta < 1,$$

где по-прежнему k означает количество действий автомата, а $\lfloor \cdot \rfloor$ означает наименьшее целое число, не мень-

шее r . В моменты времени, указываемые правилом J , строится вариационный ряд

$$V_{i_1} > V_{i_2} > \dots > V_{i_k}.$$

Для определенности здесь стоят знаки строгих неравенств. При равенстве нескольких эмпирических средних выигравшей условимся располагать соответствующие оценки по возрастанию их индексов и в таком же порядке располагать по «качеству» действия y_{i_1}, \dots, y_{i_k} .

Результат первого применения оператора O заключается в отборе группы первых («лучших») $\lfloor \theta k \rfloor$ действий $y_{i_1}, \dots, y_{i_{\lfloor \theta k \rfloor}}$ и присвоении им суммарной вероятности \hat{p} (вместо меньшей прежней вероятности $\frac{\lfloor \theta k \rfloor}{k}$), равномерно распределяемой между выбранными действиями; дополнительная вероятность $1 - \hat{p}$ поровну делится среди оставшихся. Допустим, что в момент следующего преобразования вероятностей действий отобранная группа действий сохраняет место во главе вариационного ряда

$$V_{j_1} > V_{j_2} > \dots > V_{j_k},$$

где $(j_1, \dots, j_{\lfloor \theta k \rfloor})$ — перестановка индексов $(i_1, \dots, i_{\lfloor \theta k \rfloor})$.

Тогда снова отбирается доля θ лучших действий, из числа имеющихся, $y_{j_1}, \dots, y_{j_{\lfloor \theta k \rfloor}}$ и среди них распределяется суммарная вероятность \hat{p} . Указанная процедура повторяется и приводит в конце концов к одному лучшему действию y_μ , которому приписывается вероятность \hat{p} . Распределение вероятностей действий оказывается равным

$$\mathbf{p} = \left(\frac{1-p}{k-1}, \dots, \frac{1-\hat{p}}{k-1}, \hat{p}, \frac{1-\hat{p}}{k-1}, \dots, \frac{1-\hat{p}}{k-1} \right).$$

Если в момент следующего преобразования вектора \mathbf{p} действие y_μ остается лучшим, автомат переходит в δ -режим, т. е. имеет распределение

$$\mathbf{p}_\delta = \left(\frac{\delta}{k-1}, \dots, \frac{\delta}{k-1}, 1-\delta, \frac{\delta}{k-1}, \dots, \frac{\delta}{k-1} \right).$$

Обратимся к свойствам оператора O при изменениях состава отобранной группы лучших действий. Принимаем-

мые условия зависят от того, каким классом ОПНЗ управляет автомат. В случае произвольного класса, т. е. если оптимальных действий может быть более одного, следует отбирать долю θ лучших действий независимо от того, изменился ли ее состав по сравнению с предыдущим этапом, а при реализации δ -режима переходить в δ -режим, отвечающий новому лучшему действию. В случаях, когда известно, что управляемые ОПНЗ имеют по единственному оптимальному действию, возможны кроме описанного еще иные образы поведения, например:

- переходить от текущего распределения \mathbf{p} к исходному равномерному $\mathbf{p}_0 = \left(\frac{1}{k}, \dots, \frac{1}{k}\right)$;
- переходить от имеющейся группы (из x действий) лучших действий к расширенной в θ^{-1} раз новой группе, отвечающей первым $\lceil \theta^{-1}x \rceil$ членам нового вариационного ряда;
- присваивать суммарную вероятность \hat{p} новому составу группы из x лучших действий (т. е. размер группы не меняется).

Если автомат находится в δ -режиме, то в одном варианте при смене лучшего действия вероятность его снижается с $1 - \delta$ до \hat{p} , т. е. покидается δ -режим. В другом варианте вероятность \hat{p} присваивается новому лучшему действию.

Определенное нами семейство (по $k, \theta, \hat{p}, \delta, J$) $\delta\omega$ -автоматов обозначается символом $G(k, \theta, \hat{p}, \delta, J)$, а входящие в него автоматы называются *автоматами* G .

Правило J выбора момента трансформации распределения вектора \mathbf{p} может быть инерционным или мобильным (т. е. неинерционным). К первому варианту относится, например, изменение \mathbf{p} после того, как все компоненты N возрастут по крайней мере на $n_0 \geqslant 1$. Ко второму варианту — изменение \mathbf{p} на каждом такте. При мобильном правиле J задание малого θ , но большого \hat{p} , влечет быстрое выделение небольшой группы действий, которые совершаются с большой вероятностью \hat{p} . Тогда разброс откликов ОПНЗ и из-за этого оценок V_j приводит к тому, что последние часто располагаются в порядке, отличном от истинного. Поэтому оптимальное действие совершается редко и нужно много времени для «установления истины».

Если же правило J инерционное, требующее накопления основательных, представительных сведений, то свойства автомата G слабо зависят от его параметров, ибо каждая трансформация p основана на статистических выборках значительного объема. Это дает основания для уверенности, что почти сразу будет отбираться истинная группа лучших действий.

Количественная характеристика свойств $\delta\omega$ -автоматов при управлении ими ОПНЗ может основываться на различных признаках. Например, в основу суждений об этих автоматах можно положить доход за время «обучения» (т. е. суммарный средний выигрыш до попадания в δ -оптимальный режим), среднее время первого достижения δ -оптимального режима либо окончательного попадания в этот режим и т. д. Аналитическое отыскание любой из этих характеристик в достаточно общих предпосылках встречает серьезные трудности.

В качестве $\delta\omega$ -автомата изберем автомат G . За оценку его «качества» примем математическое ожидание и дисперсию момента τ'_δ первого достижения δ -оптимального режима. Точные их выражения таковы:

$$T = E\tau'_\delta = \sum_{n \geq 0} n q_n, \quad \sigma^2 = E(\tau'_\delta - T)^2 = \sum_{n \geq 0} (n - T)^2 q_n,$$

где $q_n = P(\tau'_\delta = n)$. Называть их будем соответственно *среднее время обучения* и *дисперсия времени обучения*. Другое выражение этих характеристик имеет вид

$$T = \sum_{n=1}^{\infty} P(\tau'_\delta > n), \quad \sigma^2 = 2 \sum_{n=1}^{\infty} (n + 1) P(\tau'_\delta > n) - T^2.$$

Эта последняя запись иногда бывает удобнее.

Эти величины можно находить численно, имитацией объекта $L \otimes \xi$ на ЦВМ, в отличие от аналогичных величин, относящихся к немарковскому моменту τ''_δ .

Зададим конкретный вид автомата. Для этого нужно указать способ трансформации распределения p при изменениях состава отобранный группы лучших действий и правило J выбора моментов приложения оператора O . Правило J таково: первые $k-1$ тактов управления процессом действия совершаются с одинаковыми вероятно-

стями, а начиная с k -го такта распределение преобразуется в каждый момент времени. Действие оператора O при изменениях состава отобранный группы действий заключается в расширении ее в θ^{-1} раз; если же автомат уже находится в δ -режиме, то вероятность лучшего действия понижается с $1-\delta$ до $\hat{\rho}$.

Число действий автомата будем варьировать: $k=2^l$, $l=2, \dots, 6$; а остальные параметры примем такими: $\theta=1/2$, $\hat{\rho}=0.8$, $\delta=0.2$. Управляемый процесс описывается следующим образом. Его отклики на действие y_j суть случайные величины $\xi(j)=(1+\zeta_1/2)j+\zeta_2$, где ζ_i равномерно распределены на интервале $(-1, 1)$. Значит, средние выигрыши равны $W_j=j$ и области значений процесса, отвечающие управлению с близкими номерами, пересекаются.

Результаты расчетов в указанных предпосылках оказались такими: среднее время обучения T и среднеквадратическое уклонение σ выражаются эмпирическими формулами

$$T \approx 4k, \quad \sigma \approx 4k$$

и в пределах точности расчета совпадают. Минимальное время выхода на δ -оптимальный режим совпадает с теоретическим значением $\log_2 k + k - 1$, а максимальное время оказалось большим, так при $k=64$ оно превышает 2000 тактов. Наконец, суммарный средний выигрыш V за период обучения пропорционален k и равен $V \approx 0.66 k$.

Приведем некоторые аналитические результаты, относящиеся к автомата G , у которого за фиксированное количество тактов T_0 , зависящее от k , отбирается одно лучшее действие, и ему присваивается вероятность $1-\delta$. Начиная с этого момента времени автомат окажется в δ -режиме. Для таких автоматов мы укажем верхнюю и нижнюю оценки математических ожиданий величин τ'_δ и τ''_δ и, кроме того, в случае $k=2$ действий точное выражение среднего времени обучения.

Как и в предыдущем параграфе, будем считать множество входных сигналов автомата G лежащим на отрезке $[a, b]$.

Докажем следующее предложение.

Л е м м а 1. Пусть c_j — произвольная константа такая, что $c_j > W_j$. Тогда имеет место оценка

$$\mathbf{P} \{V_j(t) \geq c_j\} < \left(1 - \frac{\delta}{k-1} e^{-\alpha_j}\right)^t,$$

где

$$\alpha_j = 2 \left(\frac{W_j - c_j}{b - a} \right)^2, \quad j = 1, \dots, k, \quad t > T_0.$$

Доказательство. По формуле полной вероятности имеем

$$\begin{aligned} \mathbf{P} \{V_j(t) \geq c\} &= \sum_{n=0}^t \mathbf{P} \{V_j(t) \geq \\ &\geq c \mid N_j(t) - N_j(T_0) = n\} \mathbf{P} \{N_j(t) - N_j(T_0) = n\}. \end{aligned} \quad (1)$$

Вероятность $\mathbf{P} \{V_j(t) \geq c \mid N_j(t) - N_j(T_0) = n\}$ оценим с помощью леммы 1 из § 5:

$$\begin{aligned} \mathbf{P} \{V_j(t) \geq c \mid N_j(t) - N_j(T_0) = n\} &= \\ &= \sum_{m=0}^t \mathbf{P} \{V_j(t) \geq c \mid N_j(t) = n + m\} \mathbf{P} \{N_j(T_0) = \\ &= m \mid N_j(t) - N_j(T_0) = n\} = \sum_{m=0}^t \mathbf{P} \{x_j^1 + \dots + x_j^{m+n} - \\ &- \mathbf{E}(x_j^1 + \dots + x_j^{m+n}) \geq (n+m)(c_j - W_j)\} \mathbf{P} \{N_j(T_0) = \\ &= m \mid N_j(t) - N_j(T_0) = n\} \leqslant \\ &\leqslant \sum_{m=0}^t \exp \left\{ -2(n+m) \left(\frac{c_j - W_j}{b-a} \right)^2 \right\} \mathbf{P} \{N_j(T_0) = \\ &= m \mid N_j(t) - N_j(T_0) = n\} \leqslant \exp \left\{ -2n \left(\frac{c_j - W_j}{b-a} \right)^2 \right\} \times \\ &\times \sum_{m=0}^t \mathbf{P} \{N_j(T_0) = m \mid N_j(t) - N_j(T_0) = n\} = e^{-n\alpha_j}. \end{aligned} \quad (2)$$

Для оценки вероятности $\mathbf{P} \{N_j(t) - N_j(T_0) = n\}$ обозначим через $v(t)$ число лидерств величины V_j за время $[T_0 + 1, t]$

и запишем

$$\begin{aligned} \mathbb{P}\{N_j(t) - N_j(T_0) = n\} &= \\ &= \sum_{s=0}^t \mathbb{P}\{N_j(t) - N_j(T_0) = n | v_j(t) = s\} \mathbb{P}\{v_j(t) = s\}. \end{aligned}$$

Легко подсчитать вероятность события $\{N_j(t) - N_j(T_0) = n | v_j(t) = s\}$; она равна вероятности того, что действие y_j совершилось за время $[T_0+1, t]$ ровно n раз при условии, что было s возможностей совершиться с вероятностью $1-\delta$ и $t-s$ возможностей — с вероятностью $\delta/(k-1)$. Поэтому,

$$\begin{aligned} \mathbb{P}\{N_j(t) - N_j(T_0) = n | v_j(t) = s\} &= \\ &= \sum_{m=\max(0, n-t+s)}^{\min(n, s)} [C_s^m (1-\delta)^m \delta^{s-m}] \times \\ &\quad \times \left[C_{t-s}^{n-m} \left(\frac{\delta}{k-1}\right)^{n-m} \left(1 - \frac{\delta}{k-1}\right)^{t-s-n+m} \right]. \quad (3) \end{aligned}$$

В квадратных скобках стоят вероятности того, что действие y_j совершилось ровно m раз за время его лидерства и $n-m$ раз за остальное время.

Введем несколько обозначений:

$$\begin{aligned} a_j &= \frac{\delta}{k-1} e^{-\alpha_j} \left(1 - \frac{\delta}{k-1}\right), \quad b = (k-1) \left(1 - \frac{\delta}{k-1}\right) (1-\delta) \delta^{-2}, \\ d &= \frac{\delta}{1 - \frac{\delta}{k-1}}, \quad f = 1 - \frac{\delta}{k-1}. \end{aligned}$$

Из равенств (1)–(3) следует, что

$$\begin{aligned} \mathbb{P}\{V_j(t) \geq c\} &\leq \\ &\leq f^t \sum_{s=0}^t \sum_{n=0}^t a_j^n d^s \sum_{m=\max(0, n-t+s)}^{\min(n, s)} C_s^m C_{t-s}^{n-m} b^m \mathbb{P}\{v_j(t) = s\}. \quad (4) \end{aligned}$$

Имеет место тождество

$$\sum_{n=0}^t a_j^n \sum_{m=\max(0, n-t+s)}^{\min(n, s)} C_s^m C_{t-s}^{n-m} b^m = (1 + a_j)^{t-s} (1 + a_j b)^s, \quad (5)$$

в справедливости которого нетрудно убедиться, приравнивая коэффициенты при a_j^n в обеих частях.

Из (4) и (5) вытекает

$$\begin{aligned} \mathbb{P}\{V_j(t) \geq c\} &\leq f^t \sum_{s=0}^t (1+a_j)^{t-s} (1+a_j b)^s d^s \mathbb{P}\{\nu_j(t)=s\} = \\ &= [f(1+a_j)]^t \sum_{s=0}^t \left[\frac{d(1+ab)}{1+a} \right]^s \mathbb{P}\{\nu_j(t)=s\} = \\ &= \left(1 - \frac{\delta}{k-1} e^{-\alpha_j}\right)^t \sum_{s=0}^t \left[\frac{1 - (1-\delta)(1-e^{-\alpha_j})}{1 - \frac{\delta}{k-1}(1-e^{-\alpha_j})} \right]^s \mathbb{P}\{\nu_j(t)=s\}. \end{aligned}$$

Выражение в квадратных скобках меньше единицы, если $\delta < \frac{k-1}{k}$ (это условие, конечно, всегда выполнено), и так как $\sum_{s=0}^t \mathbb{P}\{\nu_j(t)=s\} = 1$, находим $\mathbb{P}\{V_j(t) \geq c\} \leq \left(1 - \frac{\delta}{k-1} e^{-\alpha_j}\right)$, что и утверждалось.

Если применить лемму 1 к величинам $-V_j$, получим, что справедлива такая лемма:

Лемма 2. Пусть c_j — произвольная константа такая, что $c_j < W_j$. Тогда выполняется неравенство

$$\mathbb{P}\{V_j(t) \leq c_j\} \leq \left(1 - \frac{\delta}{k-1} e^{-\alpha_j}\right)^t,$$

т.е.

$$\alpha_j = 2 \left(\frac{W_j - c_j}{b - a} \right)^2, \quad j = 1, \dots, k, \quad t > T_0.$$

С помощью лемм 1 и 2 выведем оценки для математического ожидания и дисперсии величин τ'_δ и τ''_δ .

Теорема 1.

$$\mathbb{E}\tau'_\delta \leq \mathbb{E}\tau''_\delta \leq \frac{k-1}{\delta} \sum_{j=1}^k \frac{1}{1 - e^{-\alpha_j}},$$

$$\mathbb{D}\tau'_\delta \leq \mathbb{D}\tau''_\delta \leq 2 \left(\frac{k-1}{\delta} \right)^2 \sum_{j=1}^k \frac{1}{(1 - e^{-\alpha_j})^2} + \frac{k-1}{\delta} \sum_{j=1}^k \frac{1}{1 - e^{-\alpha_j}},$$

где

$$\alpha_j = 2 \left(\frac{W_j - c}{b - a} \right)^2, \quad j = 1, \dots, k,$$

причем c — произвольная константа такая, что $\max_{i>1} W_i < c < W_1$.

Доказательство. Пусть $\max_{i>1} W_i < c < W_1$. Тогда

$$\begin{aligned} \mathbf{P}\{\tau''_\delta > t\} &\leq \mathbf{P}\{V_1(t) \leq \max_{i>1} V_i(t)\} \leq \\ &\leq \mathbf{P}\{V_1(t) \leq c\} + \mathbf{P}\{\max_{i>1} V_i(t) \geq c\} \leq \\ &\leq \mathbf{P}\{V_1(t) \leq c\} + \sum_{i=2}^k \mathbf{P}\{V_i(t) \geq c\}. \end{aligned}$$

Отсюда, согласно неравенствам из лемм 1, 2, следует

$$\mathbf{P}\{\tau''_\delta > t\} \leq \sum_{j=1}^k \left(1 - \frac{\alpha_j}{k-1} e^{-\alpha_j t} \right)^t. \quad (6)$$

Для целочисленной случайной величины τ''_δ справедливы равенства

$$\mathbf{E}\tau''_\delta = \sum_{t=T_0}^{\infty} \mathbf{P}\{\tau''_\delta > t\},$$

$$\mathbf{D}\tau''_\delta = 2 \sum_{t=T_0+1}^{\infty} t \mathbf{P}\{\tau''_\delta > t\} + \sum_{t=T_0}^{\infty} \mathbf{P}\{\tau''_\delta > t\} - \left(\sum_{t=T_0}^{\infty} \mathbf{P}\{\tau''_\delta > t\} \right)^2.$$

Подставляя в эти формулы оценку (6), получаем утверждение теоремы.

Займемся теперь нижней оценкой интересующих нас величин.

Теорема 2.

$$\mathbf{E}\tau''_\delta \geq \mathbf{E}\tau'_\delta \geq c \frac{k}{\delta},$$

где постоянная c не зависит от k и δ .

Доказательство. Пусть автомат функционировал до момента t следующим образом: на отрезке $[T_0 + 1, t]$ все время совершалось действие y_j ; на отрезке $[T_0 + 1, t]$

все время лидировала величина V_j ; максимальная из величин V_i , $i \neq j$, была в момент T_0 меньше, чем $W_j/2$.

Будем говорить при выполнении этих условий, что наступило событие A_j . Таким образом, в краткой записи,

$$A_j = \left\{ \max_{i \neq j} V_i(T_0) < \frac{W_j}{2}; \quad V_j(s) \geq \max_{i \neq j} V_i(s), \quad s = T_0, \dots, t; \right. \\ \left. N_j(t) - N_j(T_0) = t - T_0 \right\}.$$

Очевидно, что осуществление хотя бы одного из непересекающихся событий A_j , $j = 2, \dots, k$, влечет за собой выполнение события $\{\tau'_\delta > t\}$. Поэтому

$$\mathbf{P}\{\tau'_\delta > t\} \geq \mathbf{P}\left\{\bigcup_{j=2}^k A_j\right\} = \sum_{j=2}^k \mathbf{P}\{A_j\}. \quad (7)$$

Оценим вероятность события A_j . Пусть A_j осуществилось. Положим $M_j = \max_{i \neq j} V_i(T_0)$ и обозначим через x^1, \dots, x^{t-T_0} выигрыши, полученные на отрезке $[T_0 + 1, t]$. Учитывая, что на этом отрезке действие y_j совершалось с вероятностью $1 - \delta$, имеем

$$\mathbf{P}\{A_j\} = (1 - \delta)^{t-T_0} \mathbf{P}\left\{M_j < \frac{W_j}{2}; \quad V_j(T_0) \geq M_j; \right. \\ \left. x^1 + V_j(T_0) N_j(T_0) \geq [N_j(T_0) + 1] M_j, \dots \right. \\ \left. \dots, x^1 + \dots + x^{t-T_0} + V_j(T_0) N_j(T_0) \geq \right. \\ \geq [N_j(T_0) + t - T_0] M_j \} \geq \mathbf{P}\left\{M_j < \frac{W_j}{2}\right\} \mathbf{P}\left\{V_j(T_0) > \frac{W_j}{2}\right\} \times \\ \times \mathbf{P}\left\{x^1 + \dots + x^{t-T_0} \geq (i - T_0) \frac{W_j}{2}, \quad i = T_0 + 1, \dots, t\right\}. \quad (8)$$

Первые два сомножителя в правой части оцениваются снизу положительными константами, не зависящими от параметров k и δ . Например, для первого сомножителя

имеем

$$\mathbb{P} \left\{ M_j < \frac{W_j}{2} \right\} \geq \mathbb{P} \left\{ \text{все выигрыши на отрезке } [1, T_0] \right. \\ \left. \text{меньше } \frac{W_j}{2} \right\} \geq \min_{i \neq j} \left(\mathbb{P} \left\{ x_i < \frac{W_j}{2} \right\} \right)^{T_0} > 0,$$

где x_i — по-прежнему значение выигрыша за действие y_i . Третий сомножитель в правой части (8) обозначим через $\pi_j(t)$:

$$\pi_j(t) = \mathbb{P} \left\{ x^1 - \frac{W_j}{2} \geq 0, \dots, \sum_{n=1}^{t-T_0} \left(x^n - \frac{W_j}{2} \right) \geq 0 \right\}.$$

Из предыдущих формул следует

$$E\tau'_\delta = \sum_{t=T_0}^{\infty} \mathbb{P} \{ \tau'_\delta > t \} \geq \text{const} \sum_{j=2}^k \sum_{t=T_0}^k \pi_j(t) (1-\delta)^{t-T_0}, \quad (9)$$

причем постоянная не зависит от k и δ .

Вероятности типа $\pi_j(t)$ исследуются в теории случайных блужданий. Отметим следующий результат (Феллер, т. II, стр. 682, лемма 2):

для всех $|s| \leq 1$ справедливо тождество

$$\sum_{t=T_0}^{\infty} s^{t-T_0} \pi_j(t) = \\ = \exp \sum_{t=T_0+1}^{\infty} \frac{s^{t-T_0}}{t-T_0} \mathbb{P} \left\{ \left(x^1 - \frac{W_j}{2} \right) + \dots + \left(x^{t-T_0} - \frac{W_j}{2} \right) \geq 0 \right\}. \quad (10)$$

Очевидно, $\mathbb{P} \left\{ \left(x_j^1 - \frac{W_j}{2} \right) + \dots + \left(x_j^{t-T_0} - \frac{W_j}{2} \right) \geq 0 \right\} \geq \beta_j > 0$, где β_j не зависит от k и δ . Поэтому из (9) и (10) получаем

$$E\tau'_\delta > \min_{j>1} \beta_j (k-1) \exp \sum_{t=T_0+1}^{\infty} \frac{(1-\delta)^{t-T_0}}{t-T_0} = \text{const} \frac{k}{\delta},$$

что требовалось доказать.

Согласно формулам (7), (8)

$$\mathbb{P} \{ \tau'_\delta > t \} > \text{const} \sum_{j=2}^k (1-\delta)^t \pi_j(t).$$

Вероятность $\pi_j(t)$ убывает с ростом t не быстрее, чем экспоненциально, следовательно, то же справедливо и для вероятности $P\{\tau_\delta' > t\}$. Из этого замечания и из формулы (6) вытекает

Следствие 1. *Существуют положительные постоянные K_i , μ_i , $i=1, 2$, такие, что*

$$K_1 e^{-\mu_1 t} < P\{\tau_\delta' > t\} \leq P\{\tau_\delta'' > t\} < K_2 e^{-\mu_2 t}, \quad t > T_0.$$

Из теорем 1, 2 легко получить также следующие утверждения:

Следствие 2.

$$E\tau_\delta' = O\left(\frac{1}{\delta}\right), \quad E\tau_\delta'' = O\left(\frac{1}{\delta}\right).$$

Следствие 3. *Если существует такая положительная константа r , что выполняется неравенство $W_1 - \max_{j>1} W_j > r$, то*

$$E\tau_\delta' = O(k), \quad E\tau_\delta'' = O(k).$$

Рассмотрим простейший вариант автомата типа G , имеющего всего два действия y_1 и y_2 . Постараемся получить точные значения характеристик величины τ_δ' . Выигрыши за эти действия обозначаются через x_1 и x_2 , и пусть функции распределения F^1 и F^2 и соответственно характеристические функции φ_1 и φ_2 . Примем, что действие y_1 оптимальное, т. е. $W_1 > W_2$. Нам удобно принять, в отличие от обычного предположения о начальном состоянии, что уже в первый момент функционирования ($t=1$) автомат с вероятностью $1-\delta$ выбирает действие y_2 , т. е. он находится в δ -режиме.

Введем в рассмотрение следующие функции. Для $0 \leq m \leq t$ положим

$$\begin{aligned} G_t(w, z, m) &= \\ &= P\{V_1(t) \leq w \leq V_2(t) \leq z; \tau > t-1; N_1(t) = m\}, \\ H_t(w, z, m) &= \\ &= P\{V_2(t) \leq z < V_1(t) \leq w; \tau > t-1; N_1(t) = m\}. \end{aligned}$$

Очевидно, $G_t(w, z, m) = 0$, если $w > z$, и $H_t(w, z, m) = 0$,

если $w \leq z$. Кроме того,

$$\int_{w=-\infty}^{\infty} \int_{t=w}^{\infty} G_t(dw, dz, m) = P\{\tau > t, N_1(t) = m\},$$

$$\int_{w=-\infty}^{+\infty} \int_{t=-\infty}^{w+0} H_t(dw, dz, m) = P\{\tau = t, N_1(t) = m\}.$$

Будем считать, по определению, что G_0 есть распределение вероятностей, сосредоточенное в нуле.

Далее, положим

$$A(s, r, \mu, \nu) =$$

$$= \sum_{t=0}^{\infty} \sum_{m=0}^t s^t r^m \int_{w=-\infty}^{\infty} \int_{z=w}^{\infty} e^{i[m\mu w + (t-m)\nu z]} G_t(dw, dz, m),$$

$$B(s, r, \mu, \nu) =$$

$$= \sum_{t=1}^{\infty} \sum_{m=0}^t s^t r^m \int_{w=-\infty}^{\infty} \int_{z=-\infty}^{w+0} e^{i[m\mu w + (t-m)\nu z]} H_t(dw, dz, m).$$

Оба ряда сходятся по крайней мере при $|s| \leq 1$.

Согласно определениям этих функций

$$A(1, 1, 0, 0) = E\tau'_\delta.$$

Таким образом, для вычисления математического ожидания величины τ'_δ достаточно отыскать функцию A . Первоначально выведем тождество, связывающее функции A и B .

Л е м м а 3.

$$1 - B = \{1 - [\delta \varphi_1(\mu) r + (1 - \delta) \varphi_2(\nu)] s\} A.$$

Доказательство. Запишем рекуррентные соотношения, выражающие значения функций H_t и G_t через значение G_{t-1} . Воспользуемся наглядными вероятностными соображениями.

Пусть $0 < m < t$, $t > 1$. Предположим, что к моменту $t-1$ первое действие совершилось m раз, величины

V_1 и V_2 достигли значений u и v соответственно, а неравенство $V_1 - V_2 \leq 0$ на отрезке $[1, t-1]$ ни разу не нарушалось. Вероятность этого события равна $G_{t-1}(du, dv, m)$. В следующий момент с вероятностью δ совершаются действие y_1 и с вероятностью $1-\delta$ — действие y_2 , причем выигрыш, полученный в этот момент, может быть таким, что неравенство $V_1 - V_2 \leq 0$ нарушается. С помощью введенных обозначений описанную ситуацию можно записать следующим образом:

для $w \leq z$ справедливо равенство

$$G_t(w, z, m) =$$

$$\begin{aligned} &= \delta \int_u \int_v P \{V_1(t) \leq w \leq V_2(t) \leq z \mid V_1(t-1) = \\ &= u, V_2(t-1) = v; N_1(t) = m\} G_{t-1}(du, dv, m-1) + \\ &+ (1-\delta) \int_u \int_v P \{V_1(t) \leq w \leq V_2(t) \leq z \mid V_1(t-1) = u, \\ &V_2(t-1) = v; N_1(t) = m-1\} G_{t-1}(du, dv, m), \end{aligned} \quad (11)$$

для $z < w$ имеет место равенство

$$H_t(w, z, m) =$$

$$\begin{aligned} &= \delta \int_w \int_v P \{V_2(t) < z < V_1(t) \leq w \mid V_1(t-1) = u, \\ &V_2(t-1) = v; N_1(t) = m\} G_{t-1}(du, dv, m-1) + \\ &+ (1-\delta) \int_w \int_v P \{V_2(t) \leq z < V_1(t) \leq w \mid V_1(t-1) = u, \\ &V_2(t-1) = v; N_1(t) = m\} G_{t-1}(du, dv, m). \end{aligned} \quad (12)$$

Расставим пределы интегрирования и преобразуем правую часть в (11):

$$\begin{aligned}
 G_t(w, z, m) = & \\
 = \delta \int_{v=w}^z \int_{u=-\infty}^v & \mathbb{P} \{ V_1(t) \leqslant w \mid V_1(t-1) = u, N_1(t) = m \} \times \\
 & \times G_{t-1}(du, dv, m-1) + \\
 & + (-\delta) \int_{u=-\infty}^w \int_{v=u}^{\infty} \mathbb{P} \{ w \leqslant V_2(t) \leqslant \\
 & \leqslant z \mid V_2(t-1) = v, N_1(t) = m \} G_{t-1}(du, dv, m) = \\
 = \delta \int_{v=w}^z \int_{u=-\infty}^v & \mathbb{P} \{ x_1 \leqslant mw - (m-1)u \} G_{t-1}(du, dv, m-1) + \\
 & + (1-\delta) \int_{u=-\infty}^w \int_{v=u}^{\infty} \mathbb{P} \{ x_2 \leqslant \\
 & \leqslant (t-m)z - (t-m-1)v \} G_{t-1}(du, dv, m) - \\
 & - (1-\delta) \int_{u=-\infty}^w \int_{v=u}^{\infty} \mathbb{P} \{ x_2 < (t-m)w - (t-m-1)v \} \times \\
 & \times G_{t-1}(du, dv, m) = \\
 = \delta \int_{v=w}^z \int_{u=-\infty}^v & F^1(mw - (m-1)u) G_{t-1}(du, dv, m-1) + \\
 & + (1-\delta) \int_{u=-\infty}^w \int_{v=u}^{\infty} F^2((t-m)z - (t-m-1)v) \times \\
 & \times G_{t-1}(du, dv, m) - \\
 & - (1-\delta) \int_{u=-\infty}^w \int_{v=u}^{\infty} F^2((t-m)w - (t-m-1)v) \times \\
 & \times G_{t-1}(du, dv, m). \quad (13)
 \end{aligned}$$

Продифференцируем правую часть в (13) сначала по z ,

а затем по w . Получим

$$\begin{aligned}
 G_t(dw, dz, m) = & \\
 = m\delta \int_{u=-\infty}^z F^1(d(mw - (m-1)u)) G_{t-1}(du, dz, m-1) + & \\
 + (t-1)(1-\delta) \int_{v=w}^{\infty} F^2(d((t-m)z - (t-m-1)v)) \times & \\
 \times G_{t-1}(dw, dv, m), \quad z \geq w. \quad (14)
 \end{aligned}$$

Аналогичные операции, проделанные над равенством (12), приводят к следующей формуле, справедливой для $z < w$:

$$\begin{aligned}
 H_t(dw, dz, m) = & \\
 = m\delta \int_{u=-\infty}^z F^1(d(mw - (m-1)w)) G_{t-1}(du, dz, m-1) + & \\
 + (t-m)(1-\delta) \int_{v=w}^{\infty} F^2(d((t-m)z - (t-m-1)v)) \times & \\
 \times G_{t-1}(dw, dv, m). \quad (15)
 \end{aligned}$$

Поскольку правые части (14) и (15) совпадают, для любых z и w имеет место равенство

$$\begin{aligned}
 G_t(dw, dz, m) + H_t(dw, dz, m) = & \\
 = m\delta \int_{u=-\infty}^z F^1(d(mw - (m-1)u)) G_{t-1}(du, dz, m-1) + & \\
 + (t-m)(1-\delta) \int_{v=w}^{\infty} F^2(d((t-m)z - (t-m-1)v)) \times & \\
 \times G_{t-1}(dw, dv, m). \quad (16)
 \end{aligned}$$

Нетрудно понять, как изменится равенство (16), если $m=0$ или $m=t$. В этом случае исчезает либо первое слагаемое (если $m=0$), либо второе (если $m=t$).

Равенство (16) дает возможность получить искомое тождество. Для этого необходимо умножить обе части (16)

на $s^t r^m e^{i[m\mu w + (t-m)\nu z]}$, проинтегрировать по всей плоскости (w, z) и просуммировать по m (от 0 до t) и по t (от 1 до ∞). Выполним для примера указанные преобразования над первыми слагаемыми в правой части (16):

$$\begin{aligned} & \sum_{t=1}^{\infty} \sum_{m=0}^t s^t r^m \int_{w=-\infty}^{\infty} \int_{z=-\infty}^{\infty} m \delta \int_{u=-\infty}^z F^1(d(nw - (m-1)u)) \times \\ & \quad \times G_{t-1}(du, dz, m-1) e^{i[m\mu w - (t-m)\nu z]} = \\ & = \sum_{t=1}^{\infty} \sum_{m=0}^t \delta \int_{t=-\infty}^{\infty} \int_{u=-\infty}^z e^{i[(m-1)\mu u + (t-m)\nu z]} G_{t-1}(du, dz, m-1) \times \\ & \quad \times s^t r^m \int_{w=-\infty}^{\infty} m e^{i[m\mu w - (m-1)\mu u]} F^1(d(mw - (m-1)u)) = \\ & = \delta \varphi_1(\mu) \sum_{t=1}^{\infty} \sum_{m=0}^t s^t r^m \int_{z=-\infty}^{\infty} G_{t-1}(du, dz, m-1) \times \\ & \quad \times e^{i[(m-1)\mu u + (t-m)\nu z]} = \delta \varphi_1(\mu) r A. \end{aligned}$$

Аналогично, после преобразования второго слагаемого в правой части (16) получается $(1-\delta)\varphi_2(\nu)A$, а преобразование левой части (16) дает $-1+A+B$. Лемма доказана.

Оказывается, что полученное тождество позволяет определить функции A и B и, следовательно, $E\tau'_\delta$ и $P\{\tau'_\delta=t\}$. Для этого используем вспомогательные соотношения.

Л е м м а 4.

$$A(s, r, \mu, \nu) =$$

$$\begin{aligned} & = \exp \left\{ \sum_{t=1}^{\infty} \sum_{m=0}^t \frac{s^t}{t} r^m C_t^m \delta^m (1-\delta)^{t-m} \times \right. \\ & \quad \times \left. \int_{w=-\infty}^{\infty} \int_{(t-m)z=mw}^{\infty} e^{i(\mu w + \nu z)} F_m^1(dw) F_{t-m}^2(dz) \right\} \end{aligned}$$

и

$$B(s, r, \mu, \nu) = \exp \left\{ \sum_{t=1}^{\infty} \sum_{m=0}^t \frac{s^t}{t} r^m C_t^m \delta^m (1-\delta)^{t-m} \times \right.$$

$$\left. \times \int_{w=-\infty}^{\infty} \int_{z=-\infty}^{(t-m)z=mw} e^{i(\mu w + \nu z)} F_m^1(dw) F_{t-m}^2(dz) \right\},$$

где F_m^i означает t -кратную свертку распределения F^i .

Доказательство. Прологарифмируем тождество из леммы 3:

$$-\ln \{1 - [\delta \varphi_1(\mu) r - (1-\delta) \varphi_2(\nu)] s\} = -\ln(1-A) + \ln B.$$

Разлагая в ряд, имеем

$$\sum_{t=1}^{\infty} \frac{s^t}{t} [\delta \varphi_1(\mu) r + (1-\delta) \varphi_2(\nu)]^t =$$

$$= \sum_{n=1}^{\infty} \frac{A^n}{n} + \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(B-1)^n}{n}. \quad (17)$$

Преобразуем левую часть (17) следующим образом. Выражение в квадратных скобках возведем в степень t . Получим

$$\sum_{m=0}^t C_t^m (\delta r)^m (1-\delta)^{t-m} \varphi_1^m \varphi_2^{t-m}.$$

Пусть F_m^i означает распределение суммы t независимых случайных величин с одинаковым распределением F_i . Тогда φ_1^m есть характеристическая функция распределения F_m^i , а φ_2^{t-m} — характеристическая функция распределения F_{t-m}^2 . Произведение $\varphi_1^m \varphi_2^{t-m}$ есть характеристическая функция двумерного распределения $F_m^1(w) F_{t-m}^2(z)$, т. е.

$$\varphi_1^m(\mu) \varphi_2^{t-m}(\nu) = \int_{w=-\infty}^{\infty} \int_{z=-\infty}^{\infty} e^{i(\mu w + \nu z)} F_m^1(dw) F_{t-m}^2(dz).$$

Заменяя переменные, запишем

$$\begin{aligned} \varphi_1^m \varphi_2^{t-m} &= \int_{w=-\infty}^{\infty} \int_{z=-\infty}^{\infty} e^{i[m\mu w + (t-m)\nu z]} \times \\ &\quad \times F_m^1(d(mw)) F_{t-m}^2(d((t-m)z)) m(t-m) = \\ &= m(t-m) \int_{w=-\infty}^{\infty} \int_{z=w}^{\infty} e^{i[m\mu w + (t-m)\nu z]} \times \\ &\quad \times F_m^1(d(mw)) F_{t-m}^2(d((t-m)z)) + \\ &+ m(t-m) \int_{w=-\infty}^{\infty} \int_{z=-\infty}^{w+0} e^{i[m\mu w + (t-m)\nu z]} \times \\ &\quad \times F_m^1(d(mw)) F_{t-m}^2(d(mz)) = \Phi_m^1 + \Phi_{t-m}^2, \end{aligned}$$

причем Φ_m^1 и Φ_{t-m}^2 суть преобразования Фурье от мер, сосредоточенных соответственно на полуплоскостях ($z \geqslant w$) и ($z < w$).

Вспомним теперь, что, согласно определению, функции A и B являются суммами преобразований Фурье от мер, сосредоточенных соответственно на полуплоскостях ($z \geqslant w$) и ($z < w$). Приравнивая преобразования, сосредоточенные на одинаковых полуплоскостях, получим из (17) утверждение леммы.

Положим теперь в найденном выражении для функции A $s=r=1$, $\mu=\nu=0$. Полученный результат сформулируем в виде теоремы.

Теорема 3. Для рассматриваемого автомата типа G с двумя действиями справедлива формула

$$\mathbf{E} \tau'_\delta = \exp \left\{ \sum_{i=1}^{\infty} \sum_{m=0}^t \frac{C_t^m}{t} \delta^m (1-\delta)^{t-m} \mathbf{P} \left(\frac{x'_1 + \dots + x'_i}{m} \leqslant \frac{x'_2 + \dots + x'_{i-m}}{t-m} \right) \right\}.$$

Фигурирующие в правой части вероятности могут быть вычислены для каждого ОПНЗ. Это даст конкретное выражение момента величины τ'_δ .

§ 7. Асимптотически оптимальные автоматы

Для класса Q скалярных ОПНЗ выдвинем целью управления асимптотическую оптимальность в сильном смысле, т. е. выполнение равенства

$$P\left(\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{n=1}^t \xi_n = \bar{W}\right) = 1, \quad (1)$$

где $\bar{W} = \sup_y W(y)$. При конечном множестве Y средством достижения этой цели изберем автоматы. Можно показать, что даже при конечном фазовом пространстве управляемых процессов эти автоматы должны иметь бесконечное множество состояний.

В качестве примера асимптотически оптимального автомата рассмотрим автомат M , аналогичный M_δ из § 5.

Состояниями автомата M служат пары k -мерных векторов (N, V) . Функцию выходов определяют целое число $r \geqslant 1$ и подпоследовательность натурального ряда (n_i) . Автомат M чередует перебор по r раз подряд всех действий и совершение в течение n_i тактов лучшего действия.

Теорема 1. *Автомат M асимптотически оптимален.*

Доказательство следует из двух фактов: 1) в силу того, что все $N_i(t) \rightarrow \infty$, имеем $P\left(\lim_{t \rightarrow \infty} V_i(t) = W(y_i)\right) = 1$. Значит, начиная с некоторого момента τ_0 , оценка V_{j_0} , отвечающая оптимальному действию y_{j_0} , превысит все остальные оценки (для простоты мы считаем, что оптимальное действие единственно); 2) с момента τ_0 частота совершения действия y_{j_0} начнет расти, стремясь к 1. В выражении суммарного среднего выигрыша системы M за время t

$$V(t) = \sum_{i=1}^k \frac{N_i(t)}{t} V_i(t)$$

имеем $N_{j_0}(t)/t \rightarrow 1$, а сумма коэффициентов при остальных V_i стремится к нулю. Отсюда вытекает утверждение.

Обратимся теперь к другим подходам к синтезу асимптотически оптимальных систем. Пусть оптимальные действия ОПНЗ ξ_i имеют индексы из $I = (i_1, \dots, i_l)$. Сформу-

лируем цель управления в терминах блуждания по множеству F распределений вероятностей на Y .

Автомат A асимптотически оптимален, если для любого $\xi \in Q$

$$P\left(\lim_{t \rightarrow \infty} \sum_{j \in I} p_j(t) = 1\right) = 1. \quad (2)$$

В случае единственного оптимального действия y_{j_0} соответствующая вершина $e_{j_0} = (0, \dots, 0, 1, 0, \dots, 0)$ симплекса F служит поглощающей точкой $\overset{j_0}{\underset{\sim}{\text{блуждания}}}$, индуцированного взаимодействием автомата с управляемым процессом.

Легко видеть, что свойство (2) автомата влечет за собой справедливость (1) для каждого $\xi \in Q$.

Далее рассматриваются автоматы, состояния которых имеют вид $s = (N, V, p; \cdot)$, где векторы N, V, p имеют прежний смысл, а точка означает, что состояние может содержать дополнительные компоненты.

α -автоматом называется автомат указанной только что структуры, для которого справедливо включение с точностью до множества меры нуль

$$\{\omega: \lim_{t \rightarrow \infty} N_i(t) \rightarrow \infty, i = 1, \dots, k\} \subseteq \left\{ \omega: \lim_{t \rightarrow \infty} \sum_{j \in I} p_j(t) = 1 \right\}, \quad (3)$$

причем оба множества измеримы.

Теорема 2. Если для α -автомата $P\left(\lim_{t \rightarrow \infty} N_i(t) = \infty, i = 1, \dots, k\right) = 1$, то автомат асимптотически оптимален.

Доказательство сразу вытекает из того, что мера множества справа в (3) оказывается равной 1.

Использование этого достаточного условия в приведенном виде неудобно. Поэтому следует придать ему форму, допускающую проверку для конкретных автоматов. Применительно к рассматриваемым нами схемам это означает, например, обращение к последовательности $\bar{p}(t)$ распределений вероятностей действий автомата. Это делается с помощью такого факта (обобщенная лемма Бореля — Кантелли):

Теорема 3. Пусть $\eta_t, t \geq 1$, — последовательность случайных величин таких, что $0 \leq \eta_t \leq B < \infty$,

\mathcal{F}_t , $t \geq 0$, — неубывающая последовательность σ -алгебр на основном вероятностном пространстве и при каждом t величина η измерима относительно \mathcal{F}_t . Введем условные математические ожидания $p_t(\omega) = E(\eta_t | \mathcal{F}_{t-1})$, $t \geq 1$. Тогда ряд $\sum_{t=1}^{\infty} \eta_t(\omega)$ сходится для почти всех ω , для которых сходится $\sum_{t=1}^{\infty} p_t(\omega)$, и наоборот.

В качестве σ -алгебр \mathcal{F}_t можно принять алгебру, порожденную множествами вида $\{\omega : (\eta_1(\omega), \dots, \eta_t(\omega)) \in M\}$, где M — борелевское множество t -мерного евклидова пространства. Для наших целей удобнее считать, что \mathcal{F}_t порождено предысторией управляемого процесса, т. е. символически $\mathcal{F}_t = \sigma(\xi^{t+1}, y^t)$.

Пусть автомат A управляет ОПНЗ ξ . Положим

$$\eta_i^{(t)}(\omega) = \begin{cases} 1, & \text{если } y_t = y_i, \\ 0, & \text{если } y_t \neq y_i, \end{cases}$$

причем $\eta_i^{(t)}$ измеримо относительно \mathcal{F}_t . Тогда $N_i(t) = \sum_{l=1}^t \eta_l^{(t)}$. Выясним смысл величин p_i :

$$\begin{aligned} E(\eta_i^{(t)} | \mathcal{F}_{t-1}) &= P(y_t(\omega) = y_j | \xi^t, y^{t-1}) = \\ &= P(y_t(\omega) = y_j | s_t) = p_j(t, \omega), \end{aligned} \quad (4)$$

т. е. они равны вероятностям совершения действия y_j в момент t . В приведенной цепочке равенств второе справедливо потому, что управление процессом осуществляет автомат указанного вида и предыстория (ξ^t, y^{t-1}) однозначно определяет его очередное состояние.

Теорема 4. Выполнения равенства

$$P\left(\sum_{t=1}^{\infty} p_j(t, \omega) = \infty, j = 1, \dots, k\right) = 1$$

достаточно для асимптотической оптимальности α -автомата.

Доказательство непосредственно следует из теорем 2, 3 и равенства (4).

Обратимся к примерам α -автоматов.

Автомат A_{δ_α} построен аналогично автомatu A_{δ_ω} . Отличие заключается в том, что после того как A_{δ_α} достигнет δ -режима в момент очередной трансформации распределения p , уменьшается значение δ . Точнее, задана последовательность δ_n такая, что

$$0 < \delta_n < 1, \quad \delta_n \downarrow 0, \quad \sum_{n=1}^{\infty} \delta_n = \infty, \quad (5)$$

и в определяемые правилом J моменты времени достигнутое значение δ_n заменяется на δ_{n+1} . Лучшему действию присваивается вероятность $1 - \delta_n$, а остальным — одинаковые вероятности $\frac{\delta_n}{k-1}$.

Очевидно, автоматы A_{δ_α} являются α -автоматами.

Следствие 1. Автоматы A_{δ_α} асимптотически оптимальны.

Утверждение следует из неравенств $p_j(t) \geq \frac{\delta_t}{k-1}$, условий (5) и теоремы 4.

Автомат A_α отличается от предыдущего тем, что вероятность лучшего действия принимается равной $1 - \delta_n(t)$, где $n(t) = \min_{1 \leq j \leq k} N_j(t)$, и не требуется расходимости ряда

$\sum_{n=1}^{\infty} \delta_n$. Легко видеть, что A_α есть α -автомат.

Следствие 2. Автоматы A_α асимптотически оптимальны.

Требование, чтобы с вероятностью 1 $n(t) \rightarrow \infty$, равнозначно достаточному условию теоремы 2. Если бы $n(t)$ не стремилось к бесконечности, это означало бы, что вероятности всех действий автомата ограничены снизу положительными постоянными. Но отсюда следует, что все ряды $\sum_{j=1}^{\infty} p_j(t)$ расходятся и, значит, количества совершений $N_j(t)$ всех действий неограниченно растут. Полученное противоречие доказывает утверждение.

Автомат Z имеет множеством состояний совокупность троек (N, V, p, R) . Компонента $r_j(t)$ вектора $R(t) =$

$= (r_1, \dots, r_k)$ указывает, сколько раз за время t действие y_j было лучшим, т. е. оценка V_j оказывалась максимальной. Как в автомате G , из оценок V_i строится вариационный ряд (в моменты, диктуемые правилом J) и по известному числу $v(t)$ преобразований распределения p отбирается доля $\theta^{v(t)}$ лучших действий, которым присваиваются вероятности $\frac{1 - \delta_{r_{i_t}(t)}}{[k\theta^{v(t)}]}$, где i_t — индекс лучшего в момент t действия (величина в знаменателе принимается равной 1, если выражение в скобках оказывается меньше единицы), а остальным присваиваются вероятности

$$\frac{\delta_{r_{i_t}(t)}}{k - [k\theta^{v(t)}]}.$$

Относительно числовой последовательности δ_n предположим, что она монотонно стремится к нулю и $\sum_{n=1}^{\infty} \delta_n = \infty$.

Следствие 3. Автоматы Z асимптотически оптимальны.

В самом деле, справедливы неравенства

$$p_j(t, \omega) \geq \frac{\delta_{r_{i_t}(t)}}{k - [k\theta^{v(t)}]} \geq \frac{\delta_t}{k},$$

из которых в силу расходимости $\sum_{n=1}^{\infty} \delta_n$ и теоремы 4 вытекает утверждение.

Автомат SA задается числовой последовательностью $a(t)$, его состояния (N, V, p) . Трансформация распределения p происходит на каждом такте и имеет вид

$$p_{t+1} = p_t + a(t) [\Psi_t - p_t], \quad (6)$$

где компоненты вектора Ψ определены равенствами

$$\Psi_j = \begin{cases} 1, & \text{если } V_j = \max_i V_i, \\ 0, & \text{если } V_j < \max_i V_i. \end{cases}$$

В случае нескольких максимальных оценок лишь одна из них (например, с наименьшим номером) принимается равной 1. Элементы последовательности $a(t)$ будем считать удовлетворяющими условиям

$$0 < a(t) < 1, \quad \sum_{t=1}^{\infty} a(t) = \infty.$$

Убедимся, что автомат SA есть α -автомат. Действительно, при почти всех $\omega \in \{\omega: \lim_{t \rightarrow \infty} N_i(t) = \infty, i = \overline{1, k}\}$ эмпирические оценки $V_i(t) \rightarrow W(y_i)$ сходятся. Значит, при достаточно больших t окажется $V_{i_0}(t) > V_i(t), i \neq i_0$, где i_0 — индекс оптимального действия (вновь простоты ради считаем его единственным). При этих t преобразование (3) увеличивает вероятность p_{i_0} и уменьшает остальные вероятности. Следовательно, для $i \neq i_0$ имеем

$$p_i(t+1, \omega) = (1 - a(t)) p_i(t, \omega) = \prod_{l=t_0}^t (1 - a(l)) p_i(t_0, \omega).$$

Из условия $\sum_{l=1}^{\infty} a(l) = \infty$ следует $\prod_{l=t_0}^t (1 - a(l)) \rightarrow 0$. Значит, $\lim_{t \rightarrow \infty} p_i(t, \omega) = 0$ при $i \neq i_0$, т. е. $\lim_{t \rightarrow \infty} p_{i_0}(t, \omega) = 1$.

Будем далее считать $a(t) = \frac{1}{1+t}$.

Следствие 4. Автоматы SA асимптотически оптимальны.

В сделанном предположении рекуррентное соотношение (3) принимает вид

$$\mathbf{p}(t+1) = \frac{t}{t+1} \bar{p}(t) + \frac{1}{t+1} \Phi(t),$$

т. е. $p_i(t+1) \geq \frac{t}{t+1} p_i(t)$ (знак равенства достигается тогда и только тогда, когда $\psi_i(t) = 0$). Отсюда непосредственно вытекает расходимость рядов $\sum_{t=1}^{\infty} p_i(t)$.

Следующая конструкция не является α -автоматом.

Автомат FP имеет состояниями тройки (N, V, P) . Образуем средний выигрыш автомата за время t

$$V(t) = \frac{1}{t} \sum_{i=1}^k N_i(t) V_i(t), \quad V(0) = 0,$$

и вектор $\varphi(t) = (\varphi_1(t), \dots, \varphi_k(t))$, где

$$\varphi_i(t) = \begin{cases} V_i(t) - V(t), & \text{если } V_i(t) > V(t), \\ 0 & \text{в противном случае.} \end{cases}$$

Введем положительную величину («норму» φ) $\varphi = \sum_{j=1}^k \varphi_j$.

Распределение P на каждом шаге трансформируется:

$$P(t+1) = O P(t) = \frac{P(t) + \varphi(t)}{1 + \varphi(t)}.$$

Свойства автомата FP исследуем в случае двух действий y_1 и y_2 . Сначала будут доказаны вспомогательные утверждения.

1. Оператор O сжимающий.

В самом деле, пусть при всех достаточно больших $t \geq t_0$ сохраняется неравенство $V_1(t) - V_2(t) \geq v_0 > 0$. При таких t имеем

$$\begin{aligned} P_2(t+1) &= \frac{P_2(t)}{1 + \varphi(t)} = \frac{P_2(t)}{1 + \frac{N_2(t)}{t}[V_1(t) - V_2(t)]} \leq \\ &\leq \frac{P_2(t)}{1 + v_0 \frac{N_2(t)}{t}}. \end{aligned}$$

Итерируя неравенство T раз ($t = t_0 + T$), приходим к оценке

$$P_2(t_0 + T) \leq \frac{P_2(t_0)}{\prod_{s=0}^{T-1} \left(1 + v_0 \frac{N_2(t_0 + s)}{t_0 + s}\right)}.$$

В силу соотношения $\lim_{T \rightarrow \infty} \prod_{s=0}^{T-1} \left(1 + v_0 \frac{N_2(t_0 + s)}{t_0 + s}\right) = \infty$ имеем $P_2(t) \rightarrow 0$. Значит, вероятность лучшего действия y_1 стремится к единице.

Введем ω_0 -множества $\Lambda_i = \{\omega: \lim_{t \rightarrow \infty} N_i(t, \omega) = \infty\}$, $i = 1, 2$. Почти всюду на Λ_i $\lim_{t \rightarrow \infty} V_i(t) = W_i$.

2. Если $W_1 \neq W_2$, то $N_1(t)$ и $N_2(t)$ не могут одновременно стремиться к бесконечности.

Действительно, допустим, что пересечение $\Lambda_1 \cap \Lambda_2$ непусто. Для почти всех его точек $V_1(t) \rightarrow W_1$, $V_2(t) \rightarrow W_2$. Пусть $W_1 > W_2$, т. е. y_1 — оптимальное действие. Начиная с некоторого момента времени окажется справедливым неравенство $V_1(t) > V_2(t)$ и тогда вероятность

$p_1(t)$ начнет приближаться к 1. Исследуем ряд $\sum_{t=1}^{\infty} p_2(t)$.

Воспользуемся признаком сходимости Раабе, согласно которому ряд сходится, если величина

$$t(p_2(t)[p_2(t+1)]^{-1} - 1) = N_2(t)[V_1(t) - V_2(t)]$$

оказывается больше единицы при всех $t > t'$. Но согласно сказанному выше почти всюду на $\Lambda_1 \cap \Lambda_2$ правая часть этого равенства неограниченно растет (ибо $N_2(t) \rightarrow \infty$, а $V_1(t) - V_2(t) \rightarrow W_1 - W_2 > 0$). Поэтому ряд $\sum_{t=1}^{\infty} p_2(t)$ сходится и почти всюду на $\Lambda_1 \cap \Lambda_2$ последовательность $N_2(t)$ не стремится к бесконечности, т. е. автомат FP не может совершить каждое действие бесконечное число раз.

Из сказанного следует такой вывод: пусть множества откликов управляемого процесса одинаковы для обоих действий. Тогда положительна вероятность события: выполнены неравенства $V_1(t) < W_2$, $V_1(t) < V_2(t)$; растет вероятность $p_2(t)$ и убывает $p_1(t)$; действие y_1 совершается лишь конечное число раз. При наступлении этого события $\lim_{t \rightarrow \infty} V_2(t) = W_2$, но все значения $V_2(t)$ превосходят $V_1(t)$. Таким образом, автомат FP не является асимптотически оптимальным в классе всех ОПНЗ с одним и тем же фазовым пространством X .

Введем класс Q_2^* ОПНЗ с двумя управлениями (т. е. $Y = (y_1, y_2)$) и таких, что все отклики на одно действие больше математического ожидания откликов за другое.

Из сформулированных выше результатов заключаем: *автомат FP асимптотически оптимален в классе Q_2^** .

ГЛАВА IV

АВТОМАТЫ ДЛЯ НЕОДНОРОДНЫХ ПНЗ

§ 1. Постановка задачи

Пусть ξ_t — управляемый процесс, задаваемый меняющимися со временем вероятностями $\mu_t(M|y)$, т. е. представляет собой неоднородный процесс с независимыми значениями (ПНЗ). Средний выигрыш в момент t явно зависит от времени,

$$W_t(y) = \int_x \varphi(x) \mu_t(dx|y),$$

и доставляющее максимум — оптимальное — действие y^* меняется со временем. Оптимальная стратегия для максимизации доходов таких процессов — программная.

Нас интересуют возможности адаптивных систем при управлении классом ПНЗ с фиксированными множествами X и Y .

Очевидная трудность адаптивного подхода к управлению неоднородными процессами связана с изменением оптимального действия. Представим себе, что на временных интервалах длины t_0 действия y_1 и y_2 поочередно сменяют друг друга в качестве оптимального. Если за первый из этих интервалов система управления успела выявить оптимальное действие и частым его выбором получила значительный выигрыш, то на следующем интервале предпочтение этого действия может привести к большому ущербу, причем новое оптимальное действие, даже будучи определенным, возможно не успеет совершившись достаточно много раз. В дальнейшем убытки могут еще возрастать.

Высказанные качественные соображения дают основания полагать, что выдвигавшиеся ранее цели максимизации доходов не могут быть достигнуты для любого ПНЗ. Поэтому возникают два вопроса: 1) какие цели достига-

ются известными нам системами управления, предназначенными для ОПНЗ; 2) для каких подклассов класса всех ПНЗ с данными X и Y существуют адаптивные системы, обеспечивающие цели, близкие к тем, которые ставятся для ОПНЗ.

Содержанием дальнейших параграфов является изучение свойств автоматов из гл. III (и их модификаций) при управлении ПНЗ. Рассматриваются ПНЗ следующего вида: заданы l условных распределений $\mu^{(i)}(M|y)$ или, что то же самое, процессов типа ОПНЗ $\xi_t^{(1)}, \dots, \xi_t^{(l)}$, а также правило Π их чередования. Обозначим такие ПНЗ символом $(\mu^{(1)}, \dots, \mu^{(l)}; \Pi)$. Пространства X и Y у всех ПНЗ одинаковые.

§ 2. Конечные автоматы для бинарных ПНЗ

Рассматриваются ПНЗ ξ_t с бинарным фазовым пространством $X = \{x_1, x_2\}$ и конечным пространством управлений $Y = \{y_1, \dots, y_k\}$. Пусть $A_{k,n}$ — автомат из ϵ -оптимального семейства типа, рассмотренного в § 2 гл. III, управляет процессом ξ_t . Нас интересует величина предельного среднего выигрыша, достигаемого автоматом (как и ранее, имеется в виду, что элементам Y сопоставлены числа $x_1 \rightarrow 1, x_2 \rightarrow -1$). Ответ на этот вопрос можно получить исследованием ассоциированной марковской цепи (с доходами) (S, \mathcal{P}) , сопоставленной $A_{k,n} \otimes \epsilon$. При данном ранее определении цепи она оказывается неоднородной. Усложнением множества состояний цепь можно превратить в однородную. Чтобы сделать это, примем правило Π чередования распределений $\mu^{(i)}$ стохастическим: переход от распределения $\mu^{(i)}$ к $\mu^{(j)}$ происходит с неизменной вероятностью Δ_{ij} . Эти числа образуют матрицу $\Delta = \|\Delta_{ij}\|$, тем самым распределения образуют марковскую цепь (I, Δ) , где $I = (1, \dots, l)$ — множество индексов распределений.

Определим произведение марковских цепей (S, \mathcal{P}) и (I, Δ) как марковскую цепь $(S \times I, \mathcal{P} \otimes \Delta)$, у которой состояниями являются пары (s, i) , образованные состоянием s автомата и номером i распределения, а $\mathcal{P} \otimes \Delta$ является кронекеровским произведением матриц \mathcal{P} и Δ .

Эта цепь конечная. Естественно считать цепи (S, \mathcal{P}) и (I, Δ) эргодическими. Отсюда нетрудно вывести, что их произведение также эргодично. Это означает, что существуют предельные вероятности $\sigma_j^{(i)}$ того, что автомат совершил действие y_j и процесс ПНЗ характеризуется распределением $\mu^{(i)}$. Распределение $\mu^{(i)}$ представляет собой совокупность k пар вероятностей $q_j^{(i)}, p_j^{(i)} = 1 - q_j^{(i)}$ появления поощрения и наказания в ответ на действие y_j . Средний выигрыш за это действие равен $W_j^{(i)} = q_j^{(i)} - p_j^{(i)}$. Введенные обозначения позволяют выписать предельный средний выигрыш автомата $A_{k,n}$ при управлении ПНЗ ξ_t

$$W(A_{k,n}) = \sum_{j=1}^{l,k} W_j^{(i)} \sigma_j^{(i)}(n).$$

Перейдем к обсуждению свойств автоматов $A_{k,n}$ как систем управления ПНЗ. Сначала предположим, что средние выигрыши во всех процессах-компонентах $\xi_t^{(i)}$ упорядочены одинаково:

$$W_{j_1}^{(i)} \geq W_{j_2}^{(i)} \geq \dots \geq W_{j_k}^{(i)}, \quad i = 1, \dots, k.$$

Тогда, если автомат в ходе управления одной из компонент отыскал для нее оптимальное действие, он при смене компоненты не должен «переучиваться». Интуитивно нет сомнений, что с ростом числа n состояний автомата его предельный средний выигрыш будет стремиться к максимуму.

Теперь допустим, что оптимальные действия в процессах-компонентах не одинаковы. Ради простоты считаем $l=2$ и $k=2$ и пусть $\xi^{(1)}$ и $\xi^{(2)}$ «противоположные», это значит, что

$$W_1^{(1)} = -W_2^{(2)}, \quad W_2^{(1)} = -W_1^{(1)}.$$

Более того, примем вероятности значений процессов-компонент равными

$$\begin{aligned} p_1^{(1)} &= \frac{1-W}{2}, & p_2^{(1)} &= \frac{1+W}{2}, \\ p_1^{(2)} &= \frac{1+W}{2}, & p_2^{(2)} &= \frac{1-W}{2}, \end{aligned} \quad W > 0.$$

Стохастическую матрицу Δ вероятностей переходов изберем в виде

$$\Delta = \begin{pmatrix} 1-\Delta & \Delta \\ \Delta & 1-\Delta \end{pmatrix}, \quad \Delta < \frac{1}{2}.$$

Здесь $1/\Delta$ равно математическому ожиданию времени между сменами компонент. Итак, процесс $\xi = (\xi^{(1)}, \xi^{(2)}; \Pi)$ определен.

В качестве систем управления рассматриваются автоматы с линейной тактикой $L_{2,n}$. Марковская цепь (с доходами) $(S \times I, \mathcal{P} \otimes \Delta)$ эргодическая и имеет 4^n состояний. Предельные вероятности состояний (а с ними и действий) находятся стандартным путем, решением линейной системы уравнений. Интересующее нас выражение среднего выигрыша $W(L_{2,n})$ оказывается громоздким и не дает ясных поводов для суждения о его свойствах. Поэтому мы не станем приводить ни вывод выражения, ни его самого. Вместо этого ограничимся иллюстративным материалом: графиками зависимости предельного среднего выигрыша $W(L)$ от числа состояний, автомата n , величин W и Δ . На рис. 8 приведены такие графики для $W=1/3$ и четырех значений $\Delta=(0,001; 0,01; 0,1; 0,32)$.

Мы видим, что для каждого Δ существует наивыгоднейшая «глубина» автомата, доставляющая максимум его среднему выигрышу, который увеличивается при уменьшении Δ . Последнее обстоятельство естественно, ибо мы знаем, что средняя длительность периода однородности процесса равна $1/\Delta$: при малых Δ у автомата появляется возможность обнаружить оптимальное действие для данного процесса-компоненты и воспользоваться выгодой его применения. С уменьшением вероятности Δ для сохранения среднего выигрыша необходимо увеличивать

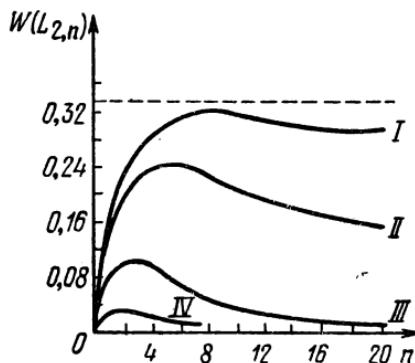


Рис. 8.

глубину n , так как переходы между процессами-компонентами осуществляются редко и автомат долгое время управляет однородным процессом. Увеличение числа n влечет монотонное стремление к нулю $W(L_{2,n})$. Причину этого нетрудно понять. При больших n за время управления одной компонентой автомат попадает в глубокие состояния той ветви, которой отвечает оптимальное действие, и когда при смене процесса на противоположный выбирается из ветви, несет крупные «убытки». Слишком маленькие n невыгодны потому, что автомат практически одинаково часто совершает оптимальное и неоптимальное действия, не задерживаясь на предпочтительной ветви.

§ 3. Оценивающие автоматы

Свойства оценивающих автоматов $O_{k,\alpha}$ (см. § 4 гл. III), как систем управления ПНЗ, мы изучим для следующего класса ПНЗ. Пусть $\xi^{(i)}, i=1, \dots, v$, — ОПНЗ с вырожденными распределениями, что, как ранее, означает, что откликом ОПНЗ $\xi^{(i)}$ на управление y_l является число x_l^i , $l=1, \dots, k$. Правило смен процессов задает стохастическая матрица $\Delta = \|\Delta_{ij}\|$, элемент Δ_{ij} которой есть вероятность замены процесса $\xi^{(i)}$ на $\xi^{(j)}$. Потребуем, чтобы ПНЗ ξ , описывался эргодической марковской цепью (I, Δ) , $I=(1, \dots, \Delta)$. Мы рассматриваем класс N таких ПНЗ.

В высказанных предположениях сформулируем цель управления автоматами $O_{k,\alpha}$ процессами из класса N . Приводимый здесь результат оправдывает введение и исследование оценивающих автоматов как адаптивных систем для ПНЗ.

Назовем *максиминным действием* y_x такое действие, что

$$\min_{1 \leq i \leq v} x_x^i = \max_{1 \leq l \leq k} \min_{1 \leq i \leq v} x_l^i.$$

Обозначим через $\pi_x(\alpha)$ предельную вероятность отвечающего действию y_x состояния автомата $O_{k,\alpha}$.

Теорема. Для любого ПНЗ из класса N предельная вероятность максиминного действия стремится к 1:

$$\lim_{\alpha \rightarrow 0} \pi_x(\alpha) = 1.$$

Доказательство. Далее предполагается, что параметр α принадлежит последовательности $\alpha_n \left(\lim_{n \rightarrow \infty} \alpha_n = 0 \right)$. Проведем вспомогательные построения. Объекту $O_k, \alpha_n \otimes \xi$ сопоставим марковскую цепь M_n , у которой состояниями являются все возможные пары $(\xi^{(i)}, s)$, где $\xi^{(i)}$ означают компоненты управляемого процесса ξ_t , а s состояние автомата O_k, α .

Пусть на множестве состояний эргодической цепи M задано разбиение $S = \{S_1, \dots, S_n\}$. Введем цепь M' , у которой переходные вероятности равны

$$p_{lm}^{M'} = \sum_{s_i \in S_l} \frac{\pi(s_i)}{\sum_{s_j \in S_m} \pi(s_j)} p_{s_i s_m}, \quad (1)$$

где $\pi(s)$ — предельные вероятности состояний цепи M , $p_{ss_m} = \sum_{s_j \in S_m} p_{ss_j}^M$. Очевидно, справедлива оценка

$$p_{lm}^{M'} \leq \max_{s \in S_l} p_{ss_m}. \quad (2)$$

Легко видеть, что предельная вероятность l -го состояния цепи M' равна сумме предельных вероятностей состояний множества S_l цепи M .

Объединим состояния $(\xi^{(1)}, s_1), \dots, (\xi^{(n)}, s_1)$ цепи M_n в множество s_l , $l = 1, \dots, k$, и рассмотрим отвечающую этому разбиению цепь M'_n . Согласно предположенной эргодичности цепи (I, Δ) для любой пары состояний $\xi^{(i)}$ и $\xi^{(j)}$ существует отличная от нуля вероятность перехода из $\xi^{(i)}$ в $\xi^{(j)}$ за d шагов ($d \leq n$). Обозначим минимальную по всем парам состояний вероятность такого перехода через γ и будем считать, что при всех l, i и достаточно больших n справедливо условие $g_{\alpha_n}(x_l^i) \leq g^r < 1$. Тогда для любого l и любой пары i, j существует $d \leq n$ такое, что вероятность перехода из состояния $(\xi^{(i)}, s_l)$ в $(\xi^{(j)}, s_l)$ цепи M_n не меньше $\gamma(1 - g)^d \geq \gamma(1 - g)^n = \gamma_1$. Это означает, что любые два состояния $s_1 \in S_l, s_2 \in S_l$, принадлежащие одному множеству, имеют в M_n одинаковые по порядку величины предельные вероятности, т. е. при всех n

$$c_1 < \frac{\pi^{M_n}(s_1)}{\pi^{M_n}(s_2)} < c_2,$$

где c_1 и c_2 — некоторые положительные константы. Отсюда в свою очередь получим, что для произвольного $s_i \in S_i$

$$\frac{\pi^{M_n}(s_i)}{\sum_{s \in S_i} \pi^{M_n}(s)} > \frac{c_1}{k}. \quad (3)$$

Легко видеть, что если $s = (\xi^{(i)}, s_l)$ и $l \neq m$, то $p_{sS_m}^{M_n} = \frac{g_{\alpha_n}(x_l^i)}{k-1}$. Введем обозначение $x_l = \min_i x_l^i$, тогда в силу (2)

$$p_{lm}^{M'_n} = \frac{g_{\alpha_n}(x_l)}{k-1}$$

и в силу (1) и (3)

$$p_{lm}^{M'_n} \geq \frac{c_1}{k(k-1)} g_{\alpha_n}(x_l),$$

т. е. *) $p_{lm}^{M'_n} \sim g_{\alpha_n}(x_l)$. Для цепи M'_n справедливо равенство

$$\frac{\pi^{M'_n}(s_l)}{\pi^{M'_n}(s_m)} = \frac{p_{ml}^{M'_n}}{p_{lm}^{M'_n}}.$$

Подставив в него найденные только что оценки, приходим к высказанному утверждению: стремиться к 1 (при $\alpha_n \rightarrow 0$) предельная вероятность действия, на котором достигается

$$\max_{1 \leq l \leq k} x_l = \max_l \min_i x_l^i$$

для любого ПНЗ из класса N . Теорема доказана.

§ 4. δ-автоматы

Если воспользоваться δ -автоматами для управления неоднородными ПНЗ, то окажется, что даже при очень медленном изменении свойств процесса автомат может получать выигрыш, много меньший максимально возможного. Причиною этого служит медленное преобразование

*) Будем писать $x_n \sim y_n$, если при некоторых $c_1, c_2 > 0$ для всех n выполнено $c_1 \leq \frac{x_n}{y_n} \leq c_2$.

распределения ρ , связанное с тем, что в основе преобразований лежат оценки средних выигрышей за действия. Оценки представляют собой средние арифметические откликов процесса и обладают двумя важными для нас свойствами: ранние и поздние значения процесса входят симметрично (т. е. равноправно) и они устойчивы.

Проиллюстрируем сказанное численным примером. В качестве $\delta\omega$ -автомата выберем тот вариант автомата G , который фигурировал в примере § 6 гл. III. В качестве управляемого им неоднородного ПНЗ изберем процесс $\xi = (\xi^{(1)}, \xi^{(2)}; \Pi)$, где $\xi^{(1)}$ — ОПНЗ из того же примера, $\xi^{(2)} = -\xi^{(1)}$, а правило Π заменяет одну компоненту на другую после того, как автомат G впервые окажется в δ -оптимальном режиме для данной компоненты. Нас интересует величина среднего времени «переобучения», т. е. математическое ожидание того времени, которое затратит автомат в ходе управления новой компонентой до попадания в δ -оптимальный режим этой компоненты. Таким образом, процессы — компоненты выбраны противоположными, оптимальное действие для одного из них является худшим для другого. Аналитические методы успеха пока не дают, и приходится ограничиваться моделированием объекта $G \otimes \xi$ на ЦВМ.

Результаты вычислений в этом примере оказались следующими: среднее время переобучения в 2,5 — 4 раза больше среднего времени обучения. Кроме того, среднее время переобучения и среднеквадратическое уклонение этого времени совпадают (в пределах точности расчета). Средний выигрыш G за время переобучения пропорционален числу действий. Численные значения указаны ниже.

Можно сделать вывод, что автоматы G неэффективны при управлении неоднородными ПНЗ, так как требуется значительное время для приспособления к меняющимся свойствам процесса и они не успевают, вообще говоря, отыскивать и преимущественно использовать текущее оптимальное действие.

Постараемся так видоизменить $\delta\omega$ -автоматы, чтобы сократить время переобучения.

Откажемся от требования состоятельности оценок $V_i(t)$ средних выигрышей за действия. Во всем остальном

будем использовать прежнюю схему δ-автоматов. Смысл компонент вектора N может изменяться (не быть равным количествам совершений действий за время t). Охарактеризованные так автоматы назовем δ -автоматами и обозначим A_δ .

Для δ-автоматов теряют силу результаты § 5 гл. III, касающиеся δ-оптимальности при управлении ОПНЗ. Разумеется, такие автоматы способны оказаться в δ-оптимальном режиме, особенно если средние выигрыши $W(y_j)$ достаточно разнесены. Однако можно быть уверенным в общем случае, что δ-автомат бесчисленное число раз покинет δ-оптимальный режим.

Задание δ-автоматов должно включать в себя операторы O_N и O_V , трансформирующие в каждый момент времени векторы N и V .

Содержанием оставшейся части этого параграфа является изложение свойств δ-автоматов, предназначенных для управления неоднородными ПНЗ.

Автоматами A_{δ^3} называются такие δ-автоматы, у которых оценки $V_j(t)$ средних выигрышей вычисляются по формуле

$$V_j^{(\beta)}(N_j) = \frac{\sum_{l=1}^{N_j} x_l^{(j)} \beta^{N_j-l}}{\sum_{l=1}^{N_{j-1}} \beta^l}, \quad 0 < \beta < 1, \quad j = 1, \dots, k,$$

или

$$V_j^{(\beta)}(N_j) = \frac{1 - \beta}{1 - \beta^{N_j}} \sum_{l=1}^{N_j} x_l^{(j)} \beta^{N_j-l}. \quad (1)$$

Здесь $x_1^{(j)}, \dots, x_{N_j}^{(j)}$ — последовательность значений управляемого процесса при управлении y_j . Формулой (1) задан оператор O_V , а O_N имеет вид

$$O_N N_j = \begin{cases} N_j + 1, & \text{если совершено } y_j, \\ N_j, & \text{в противном случае.} \end{cases}$$

Параметр β назовем *коэффициентом забывания*. Ясно, что вклад начальных значений процесса в оценку V_j убывает с экспоненциальной скоростью при $N_j \rightarrow \infty$.

$\delta \omega$ -автоматы можно интерпретировать как «предел» автоматов $A_{\delta\beta}$ при $\beta \rightarrow 1$.

Займемся свойствами автоматов $A_{\delta\beta}$ при управлении ОПНЗ. Прежде всего, оценки $V_j^{(\beta)}$ несмещенные:

$$\mathbf{E}V_j^{(\beta)} = W(y_i), \quad j = 1, \dots, k.$$

Обозначая через σ_j^2 дисперсию ОПНЗ при управлении y_j , находим дисперсию оценок

$$\bullet \quad \mathbf{D}V_j^{(\beta)}(N_j) = \frac{1 + \beta^{N_j}}{1 - \beta^{N_j}} \frac{1 - \beta}{1 + \beta} \sigma_j^2.$$

Предельная дисперсия оценок равна

$$D_j^{(\beta)} = \lim_{N \rightarrow \infty} \mathbf{D}V_j^{(\beta)}(N) = \frac{1 - \beta}{1 + \beta} \sigma_j^2,$$

т. е. хотя и меньше, чем σ_j^2 , но положительна. Отсюда вытекает, что оценка $V_j^{(\beta)}$ не состоятельная. Для уменьшения дисперсии следовало бы выбирать близкие к 1 значения β . Но при малых N это не дает выгоды из-за присутствия дополнительного множителя $\frac{1 + \beta^N}{1 - \beta^N}$.

Переходя к особенностям $A_{\delta\beta}$ при управлении неоднородными ПНЗ, допустим, что на временном интервале $(0, t)$ автомат управляет ОПНЗ $\xi_t^{(1)}$ (относящиеся к нему характеристики снабдим индексом 1). С момента $t+1$ автомат управляет ОПНЗ $\xi_t^{(2)}$ (его характеристики имеют индекс 2). Пусть на первом этапе — управляя $\xi_t^{(1)}$ — автомат совершил $N^{(1)}$ раз действие y , а управляя ОПНЗ $\xi_t^{(2)}$ — совершил $N^{(2)}$ раз. Тогда значение оценки $V_y^{(\beta)}$ равно

$$V^{(\beta)}(N^{(1)} + N^{(2)}) = \frac{\left(\sum_{l=1}^{N^{(1)}} x_l^{(1)} \beta^{N^{(1)}-l} \right) \beta^{N^{(2)}} + \sum_{l=1}^{N^{(2)}} x_l^{(2)} \beta^{N^{(2)}-l}}{1 + \beta + \dots + \beta^{N^{(1)}} + \dots + \beta^{N^{(1)}+N^{(2)}-1}} = \\ = \frac{1 - \beta}{1 - \beta^{N^{(1)}+N^{(2)}}} \left[\left(\sum_{l=1}^{N^{(1)}} x_l^{(1)} \beta^{N^{(1)}-l} \right) \beta^{N^{(2)}} + \sum_{l=1}^{N^{(2)}} x_l^{(2)} \beta^{N^{(2)}-l} \right].$$

Отсюда видно, что заключенный в сумме $\sum_{l=1}^{N^{(1)}} x_l^{(1)} \beta^{N^{(1)}-l}$

«опыт» управления первым процессом «забывается» с экспоненциальной скоростью. Первые два момента оценки имеют вид

$$\mathbb{E} V_y^{(\beta)} (N^{(1)} + N^{(2)}) = \frac{W_y^{(1)} \beta^{N^{(2)}} (1 - \beta^{N^{(1)}}) + W_y^{(2)} (1 - \beta^{N^{(2)}})}{1 - \beta^{N^{(1)}+N^{(2)}}},$$

$$\mathbb{D} V_y^{(\beta)} (N^{(1)} + N^{(2)}) = \frac{1 - \beta^{\sigma^{(1)^2}} \beta^{2N^{(2)}} (1 - \beta^{2N^{(1)}}) + \sigma^{(2)^2} (1 - \beta^{2N^{(2)}})}{(1 - \beta^{N^{(1)}+N^{(2)}})^2}.$$

Они указывают, сколь быстро происходит «стирание из памяти» автомата следов управления процессом $\xi_t^{(1)}$ после замены его на $\xi_t^{(2)}$ и длительность промежутка времени, в течение которого должно сохраняться $\xi_t^{(2)}$, чтобы автомат сумел обнаружить новое оптимальное действие. Из таких рассмотрений можно сделать вывод (впрочем, ясный из общих соображений), что управление неоднородными ПНЗ с быстрой сменой участков однородности не может быть эффективным: отсутствует не только свойство δ-оптимальности, но и целесообразность управления. Последнее означает, что средний выигрыш автомата должен удовлетворять начиная с некоторого τ неравенству

$$V(t) > \frac{1}{k} \sum_{i=1}^k W(y_i).$$

Перейдем к другим способам оценивания средних выигрышей за действия. Теперь они будут опираться на модификации средних арифметических. Рассматриваемые ниже автоматы, именуемые δh -автоматами, определяются целым числом $h \geqslant 1$ и способом Φ подсчета оценок. Забегая вперед, заметим, что $\delta\omega$ -автоматы можно трактовать как «предел» δh -автоматов при $h \rightarrow \infty$. Число h естественно называть *параметром памяти автомата*.

Перечислим некоторые из способов Φ .

Способ $\Phi_{1,a}$. Временная ось t разбивается на примыкающие отрезки длины h : $[1, h]$, $[h+1, 2h]$,

$[2h+1, 3h], \dots$. На каждом из отрезков $[(\alpha-1)h+1, \alpha h]$ оценка V_j подсчитывается как среднее арифметическое откликов процесса на j -е действие. На следующем временном отрезке $[\alpha h+1, (\alpha+1)h]$ все оценки вычисляются заново, причем до первого совершения на нем действия y_j оценка V_j остается той же, что на предыдущем от-

резке, т. е. $V_j^{(\alpha-1)} = \frac{1}{N_j^{(\alpha-1)}} \sum_{l=1}^{N_j^{(\alpha-1)}} x_l^{(j)}$. Здесь $N_j^{(\alpha-1)}$ — количество выборов автоматом j -го управления на $(\alpha-1)$ -м отрезке и подчинено ограничению $N_j^{(\alpha-1)} \leq h$ (отсюда легко уяснить вид оператора O_N). Полученная этим способом оценка несмещенная и несостоятельная при управлении ОПНЗ.

Способ $\Phi_{1,b}$. Оценка выигрыша за j -е действие представляет собой среднее арифметическое откликов процесса в течение h совершений этого действия. При следующем $(h+1)$ -м совершении его оценка вычисляется заново. Соответствующий оператор O_N действует на вектор N так:

$$O_N N_j = \begin{cases} N_j + 1 \pmod{h}, & \text{если совершено } y_j, \\ N_j & \text{в противном случае.} \end{cases}$$

Получаемые оценки снова несмещенные и несостоятельные при управлении ОПНЗ.

Достоинство указанных способов состоит в полном игнорировании прошлого, отдаленного либо h тактами времени, либо h совершениями действия. Это позволяет автомату достаточно быстро выявлять изменение свойств управляемого процесса. Накопленный автоматом «опыт» частично может сохраняться в распределении вероятности действий. Если правило J требует не слишком частого преобразования распределения, то следы прежних оптимальных или предпочтительных действий хранятся в вероятностях совершения действий. Заметим, однако, что по смыслу δh -автоматов правило J должно быть мобильным, допуская за время порядка h переход из одного δ -оптимального режима в другой. Недостаток способа заключается в невозможности со-

хранить δ-оптимальный режим в интервале однородности процесса.

Следующие варианты способов Φ более инерционны и хранят информацию об эффективности давно совершенных действий, но со временем «удельный вес» этих сведений убывает.

Способ $\Phi_{2,a}$ сходен с $\Phi_{1,a}$ и предполагает разбиение числовой прямой на отрезки $[(\alpha-1)h+1, \alpha h]$, $\alpha = 1, 2, \dots$. Из полученных на j -е действие откликов $x^{(j)}$ на первом отрезке формируется среднее

$${}_hV_j(N) = \frac{1}{N} \sum_{l=1}^N x_l^{(j)}, \quad N \leq h,$$

где N — число совершенений действия y_j на отрезке $[1, h]$. К моменту $t=h+1$ имеем оценки ${}_hV_1^{(1)}, \dots, {}_hV_k^{(1)}$, некоторые из которых — нули, так как соответствующие действия ни разу не совершились. На следующем временном отрезке оценки вычисляют по формуле

$${}_hV_j(N+l) = \frac{{}_hV_j^{(1)} + x_{N+l}^{(j)} + \dots + x_{N+l}^{(j)}}{l+1}.$$

Если N' — количество совершенений y_j на этом отрезке, то на третьем отрезке оценка равна

$${}_hV_j(N+N'+l) = \frac{{}_hV_j^{(2)} + x_{N+N'+1}^{(j)} + \dots + x_{N+N'+l}^{(j)}}{l+1},$$

где обозначено ${}_hV_j^{(2)} = {}_hV_j(N+N')$. Далее поступаем аналогичным образом.

Способ $\Phi_{2,b}$. В течение первых h совершенений j -го действия оценка ${}_hV_j$ представляет собой среднее арифметическое получаемых от процесса откликов и после h -го его совершения имеем

$${}_hV_j^{(1)} = \frac{1}{h} \sum_{l=1}^h x_l^{(j)}.$$

При дальнейших выборах этого действия полагаем

$${}^hV_j(h+l) = \frac{{}^hV_j^{(1)} + x_{h+1} + \dots + x_{h+l}}{l+1}, \quad l \leq h.$$

Введем для краткости ${}^hV_j^{(2)} = \frac{1}{h+1} \left({}^hV_j^{(1)} + \sum_{l=1}^h x_l^{(j)} \right)$ и будем вычислять оценку по формуле

$${}^hV_j(2h+l) = \frac{{}^hV_j^{(2)} + x_{2h+1} + \dots + x_{2h+l}}{l+1}, \quad l \leq h.$$

Вообще, обозначая ${}^hV_j^{(\alpha)} = {}^hV_j(\alpha h)$, имеем

$${}^hV_j(\alpha h+l) = \frac{{}^hV_j^{(\alpha)} + x_{\alpha h+1}^{(j)} + \dots + x_{\alpha h+l}^{(j)}}{l+1}, \quad l \leq h.$$

Нетрудно видеть, что оценки, получаемые способами $\Phi_{2,a}$ и $\Phi_{2,b}$, являются несмешенными и несостоительными при управлении ОПНЗ. Нас интересуют свойства δh -автоматов, управляющих ПНЗ. Пусть $\xi = (\xi^{(1)}, \xi^{(2)}; \Pi)$ — такой процесс, причем компонента $\xi^{(1)}$ заменилась на $\xi^{(2)}$ после того, как y_j было совершено αh раз ($\alpha \geq 1$) в ходе управления процессом $\xi^{(1)}$, а на следующем отрезке однородности совершилось $\beta h \geq h$ раз. В этих допущениях математическое ожидание и дисперсия оценок, получаемых способом $\Phi_{2,b}$, равны соответственно

$$\begin{aligned} \mathbb{E}^h V_j(\alpha h + \beta h) &= W_j^{(1)} (1+h)^{-\beta} + W_j^{(2)} (1 - (1+h)^{-\beta}), \\ D^h V_j(\alpha h + \beta h) &= \left(1 + \frac{2}{h(1+h)^{2(\alpha-1)}}\right) \frac{(\sigma_j^{(1)})^2}{(h+2)(1+h)^{2\beta}} + \\ &\quad + (1 - (1+h)^{-2\beta}) \frac{(\sigma_j^{(2)})^2}{h+2}. \end{aligned}$$

Рассмотрение этих выражений приводит к таким же выводам относительно δh -автоматов, которые были ранее сделаны применительно к $\delta\beta$ -автоматам. Отметим еще некоторые особенности. Для определенности будем говорить о способе $\Phi_{2,b}$.

После замены ОПНЗ $\xi^{(1)}$ на ОПНЗ $\xi^{(2)}$ происходит экспоненциально быстрое «стирание опыта», накоплен-

ного автоматом за время управления $\xi^{(1)}$. Основанием показательной функции служит число $(1+h)^{-1}$, которое меньше $1/2$ при $h > 1$. Далее, на отрезках однородности управляемого процесса оценки подсчитываются в виде средних арифметических откликов на серии h совершений действий. Такие оценки во многих случаях почти наилучшие (иногда даже эффективные) по критериям математической статистики. Наконец, δh -автоматы имеют конечное множество значений векторов N . Иногда это приводит к конечности ассоциированной марковской цепи.

Укажем примеры δ -автоматов. Пусть оператор O , действующий на распределение вероятностей действий, имеет тот же вид, что у автомата G . Это значит, что в соответствующие моменты времени строится вариационный ряд, отбирается группа лучших действий, которым передается большая вероятность. Автоматы с такой функцией переходов называются *автоматами типа G*. Если в обозначении такого автомата необходимо указать использованный метод вычислений оценок, например, так, как у $\delta\beta$ - или δh -автоматов, будем писать $G(\beta)$ или $G(h)$.

Нам следует теперь избрать критерий оценки возможностей δh -автоматов как средства управления неоднородными ПНЗ. В качестве его примем близость к единице отношения средних времен обучения T_o и переобучения T_n . Естественность такого выбора очевидна, однако к трудностям вычисления T_o добавляет не меньшие трудности отыскания T_n . Мы не будем вдаваться в аналитические тонкости, а приведем лишь некоторые численные примеры, наглядно демонстрирующие особенности автоматов.

Рассматриваются автоматы типа $G(h)$, у которых оценки находятся способом $\Phi_{2,a}$. В остальном автоматы идентичны автоматам G из примера § 6 гл. III и начала этого параграфа, включая численные значения параметров. Управляемым процессом ξ_t служит неоднородный ПНЗ, который также фигурировал в этом параграфе.

В табл. 4.1 приведены соответственно в числителе и знаменателе каждой клетки значения T_n и σ_n (среднеквадратическое уклонение времени переобучения), полученные после 100-кратной имитации объекта $G(h) \otimes \xi$. После выхода автомата на δ -оптимальный режим по отношению

Таблица 4.1

$k \backslash h$	10	15	50	∞
4	14	16	19	46
	12	13	16	32
8	33	31	42	67
	35	30	42	65
16	90	87	96	180
	84	82	91	160
32	180	170	200	330
	170	190	220	300

к компоненте $\xi^{(1)}$ процесса ξ , эту компоненту заменяет $\xi^{(2)}$. Вычисляется среднее время, прошедшее с момента замены до выхода автомата на новый δ -оптимальный режим. Во всех проведенных испытаниях автомат на него выходил. Последний столбец относится к автоматам G (т. е. предельному случаю $h=\infty$).

Из данных таблицы можно сделать такие выводы:

1. Среднее время T_n является монотонно возрастающей функцией h . При выбранных h оказывается, что T_n практически совпадает с T_0 -средним временем обучения.

2. Величины T_n и σ_n совпадают (в пределах точности расчета).

3. Среднее время T_n пропорционально числу действий k .

Таким образом, судя по этому примеру, δh -автоматы при подходящих — малых — значениях параметра памяти h оказались приемлемым средством управления неоднородными ПНЗ.

В заключение обратим внимание на то обстоятельство, что изложенные в этом параграфе алгоритмы управления действенны не только для процессов с конечным множеством значений. Очевидно, что фазовым пространством неоднородных ПНЗ может служить вся числовая прямая. Требуется соблюдение условия о существовании средних выигрышей.

§ 5. Добавления

Здесь мы рассмотрим несколько задач, отличающихся от тех, что изучались в этой главе, но решаемых автоматами из ϵ -оптимальных семейств.

Первая из задач относится к прогнозу и заключается в следующем. Задан однородный l -связный марковский процесс ξ_t с фазовым пространством $X=\{0, 1\}$ из двух элементов. Его переходные вероятности $p(\xi_t | \xi_{t-l}, \xi_{t-l+1}, \dots, \xi_{t-1})$. Будем предполагать что существуют предельные вероятности $\pi_\lambda = \lim_{t \rightarrow \infty} p(\xi_{t-1} = \lambda_1, \dots, \xi_{t-l} = \lambda_l)$, где $\lambda = (\lambda_1, \dots, \lambda_l)$ — l -мерный набор из нулей и единиц.

Требуется наилучшим образом прогнозировать будущее течение процесса. Это означает, что нужно построить такой функционал φ на траекториях процесса, чтобы его значение $\hat{\xi}_t = \varphi(\xi^{t-1})$ было оптимальной, с точки зрения некоторого критерия, оценкой величины ξ_t . Из l -связности и бинарности процесса заключаем, что функционал искать в виде булевской функции l аргументов, т. е. считаем

$$\hat{\xi}_t = \varphi(\xi_{t-1}, \dots, \xi_{t-l}).$$

Ошибку прогноза полагаем равной $\Delta_t = \xi_t - \hat{\xi}_t$ (сумма по модулю 2), она равна 1 тогда и только тогда, когда $\xi_t \neq \hat{\xi}_t$. Оптимальность прогноза означает минимальность предельной средней ошибки прогноза.

Положим сначала, что известны вероятностные характеристики прогнозируемого процесса. Вычислению предельной средней ошибки предшествует обозначения, принятые в теории булевых функций, ξ — отрицание, т. е.

$$\xi = \begin{cases} 1, & \text{если } \xi = 0, \\ 0, & \text{если } \xi = 1, \end{cases}$$

ξ^λ означает

$$\xi^\lambda = \begin{cases} \xi, & \text{если } \lambda = 1, \\ \xi, & \text{если } \lambda = 0, \end{cases}$$

наконец, при $\lambda = (\lambda_1, \dots, \lambda_l)$ принимаем $\xi_t^\lambda = \xi_{t-1}^{\lambda_1} \dots \xi_{t-l}^{\lambda_l}$. Если прогноз осуществляется с помощью функции φ , то средняя ошибка прогноза равна

$$E\Delta_t = \sum_\lambda E(\Delta_t^\varphi | \xi_t^\lambda = 1) p(\xi_t^\lambda = 1) = \sum_\lambda p(\Delta_t^\varphi = 1 | \xi_t^\lambda = 1) p(\xi_t^\lambda = 1).$$

Ясно, что $\xi_t^\lambda = 1$ лишь при $\xi_{t-1} = \lambda_1, \dots, \xi_{t-l} = \lambda_l$.

Введем еще одно обозначение

$$p_\lambda = \min(p(1 | \lambda_1, \dots, \lambda_l), p(0 | \lambda_1, \dots, \lambda_l)).$$

Предельная средняя ошибка прогноза оценивается снизу:

$$\lim_{t \rightarrow \infty} E\Delta_t^\varphi = \sum_{\lambda} \lim_{t \rightarrow \infty} p(\xi_t \neq \varphi(\xi_{t-1}, \dots, \xi_{t-l}) | \zeta_t^\lambda = 1) \pi_\lambda \geq \sum_{\lambda} p_\lambda \pi_\lambda.$$

Наименьшую ошибку обеспечивает метод максимального правдоподобия, который приводит к следующему виду прогнозирующего функционала:

$$\varphi(\lambda_1, \dots, \lambda_l) = \begin{cases} 1 & \text{при } p(1 | \lambda) \geq p(0 | \lambda), \\ 0 & \text{при } p(1 | \lambda) < p(0 | \lambda). \end{cases}$$

Ему отвечает

$$\min_{\varphi} \lim_{t \rightarrow \infty} E\Delta_t^\varphi = \sum_{\lambda} p_\lambda \pi_\lambda,$$

где минимум берется по всем булевским функциям l аргументов.

Перейдем теперь к «адаптивному» прогнозу. Он состоит в синтезе прогнозирующей системы для класса процессов, а не для конкретного полностью заданного процесса. Иными словами, функционал-оценка $\varphi(\xi_{t-1}, \dots, \xi_{t-l})$ должен быть подобран в процессе взаимодействия системы и процесса, а не задан заранее. Произвольная булевская функция единственным образом изображается совершенной дизъюнктивной нормальной формой

$$\varphi(\xi_{t-1}, \dots, \xi_{t-l}) = \sum_{\substack{(\lambda_1, \dots, \lambda_l) \\ \varphi(\lambda_1, \dots, \lambda_l)=1}} V_{\lambda_1 \dots \lambda_l} \xi_{t-1}^{\lambda_1} \dots \xi_{t-l}^{\lambda_l} = \sum_{\lambda : \varphi(\lambda)=1} V_{\lambda} \xi_t^{\lambda},$$

в которой коэффициенты y_λ равны либо 1, либо 0. В обычной постановке задачи прогноза эти коэффициенты находятся заранее, например, методом максимального правдоподобия. В адаптивной ситуации они подбираются в ходе взаимодействия прогнозирующей системы и процесса и так, чтобы минимизировать предельную среднюю ошибку прогноза. Более точная формулировка заключается в требовании построить семейство прогнозирующих систем (Π_n) , причем предельная средняя ошибка прогноза, осуществляемого системой Π_n , отличается от минимальной ошибки менее чем на ϵ_n , где $\lim_{n \rightarrow \infty} \epsilon_n = 0$. Укажем конструкцию системы Π_n .

Рассматриваются автоматы $A_{2,n}^{(\lambda_1, \dots, \lambda_l)}$ из ϵ -оптимальных семейств для бинарных ОПНЗ. Каждый из автоматов имеет два действия: одно приписывает коэффициенту $y_{\lambda_1, \dots, \lambda_l}$ значение 1, а другое — 0. Автомат $A_{2,n}^{(\lambda_1, \dots, \lambda_l)}$ совершает действие и смену состояния после наступления события $\zeta_t^{(\lambda_1, \dots, \lambda_l)}=1$ и получения на входе сигнала ошибки Δ_t . Величина $\Delta_t=0$ означает поощрение, а $\Delta_t=1$ — наказание. Всего в систему Π_n входит 2^l автоматов, из которых в каждый момент времени совершает действие в точности один, воздействующий лишь на один коэффициент дизъюнктивной формы φ . После совершения автоматом действия «единица» он наказывается с вероятностью $p(1 | \xi_{t-1})$

$= \lambda_1, \dots, \xi_{t-l} = \lambda_l$), а после действия «нуль» — с вероятностью $p(\xi_t = 0 | \xi_{t-1} = \lambda_1, \dots, \xi_{t-l} = \lambda_l)$. Отсюда следует, что каждый автомат $A_2^{(\lambda_1, \dots, \lambda_l)}$ взаимодействует с ОПНЗ, порождаемым функционалом φ и прогнозируемым марковским процессом. Этот автомат «включается» при выполнении $\zeta_t^{(\lambda_1, \dots, \lambda_l)} = 1$, которое с вероятностью 1 наступает бесконечное число раз, если $\pi_{\lambda_1, \dots, \lambda_l} > 0$. Обозначим через $m_n^{(\lambda_1, \dots, \lambda_l)}(t)$ математическое ожидание наказания автомата $A_2^{(\lambda_1, \dots, \lambda_l)}$ в момент t функционирования (т. е. при условии $\zeta_t^{(\lambda_1, \dots, \lambda_l)} = 1$). Так как автоматы образуют ϵ -оптимальное семейство, выполнено предельное равенство

$$\lim_{n \rightarrow \infty} \lim_{t \rightarrow \infty} m_n^{(\lambda_1, \dots, \lambda_l)}(t) = p_{\lambda_1, \dots, \lambda_l}.$$

Свойства семейства прогнозирующих систем Π_n по отношению к классу M_2 бинарных l -связных однородных марковских процессов, у которых существует предельное распределение, указывает следующая теорема.

Теорема 1. Справедливо равенство

$$\lim_{n \rightarrow \infty} E\Delta_t^{\Pi_n} = \min_{\varphi} \lim_{t \rightarrow \infty} E\Delta_t^{\varphi}.$$

Доказательство. Средняя ошибка прогноза системы Π_n в момент t равна

$$E\Delta_t^{\Pi_n} = \sum_{\lambda} E(\Delta_t^{\Pi_n} | \zeta_t^{\lambda} = 1) P(\zeta_t^{\lambda} = 1) = \sum_{\lambda} m_n^{\lambda}(t) P(\zeta_t^{\lambda} = 1).$$

Из приведенных выше соотношений и существования предельного распределения получаем

$$\lim_{n \rightarrow \infty} \lim_{t \rightarrow \infty} E\Delta_t^{\Pi_n} = \sum_{\lambda} p_{\lambda} \pi_{\lambda},$$

которое означает доказываемое утверждение.

Итак, прогнозирующие адаптивные системы Π_n образуют в классе процессов M_2 ϵ -оптимальное семейство в том смысле, что для любого $\epsilon > 0$ найдется такое целое n_{ϵ} , что при всех $n > n_{\epsilon}$ и достаточно больших t справедливо неравенство для средней ошибки прогноза

$$E\Delta_t^{\Pi_n} > \min_{\varphi} \min_{t \rightarrow \infty} E\Delta_t^{\varphi} - \epsilon,$$

каков бы ни был процесс из класса M_2 .

Следующая задача также относится к прогнозированию эволюции марковских процессов, но в иной постановке.

Задана конечная однородная марковская цепь (S, \mathcal{P}) , течение которой надо прогнозировать, минимизируя средние риски. Последнее понятие означает следующее: если текущее состояние цепи s_i ,

указанное прогнозом s_j , и фактически наступившее s_l , то терпится убыток ξ_{ijl} — случайная величина с распределением F_{ijl} . Средним риском в состоянии s_i называется функция на цепи $R_i(j) = \sum_l p_{il} E\xi_{ijl}$, $j = 1, \dots, |S|$. Если матрица \mathcal{P} и распределения F_{ijl} известны, алгоритм прогнозирования очевиден: после состояния s_i предполагается наступающим состояние s_{j_0} , где j_0 — индекс минимального риска $R_i(j_0) = \min_j R_i(j)$.

В адаптивной ситуации, когда \mathcal{P} и F_{ijl} неизвестны, воспользуемся $|S|$ автоматами, за которые примем $\delta\omega$ -автоматы, например автоматы G . Сопоставим каждому состоянию s_i цепи свой автомат G_i . Его выходными сигналами служат предсказываемые состояния цепи, а входными — убытки за ошибки прогноза. Легко понять, что совокупность G_i δ -оптимальных автоматов минимизирует с точностью до ϵ (определенного величиной δ) средние риски прогнозирования.

Третьей интересующей нас задачей является задача типа условного экстремума.

На траекториях ОПНЗ ξ_t с фазовым пространством X определены функционалы $\varphi_t = \varphi(\xi_t)$ и $\psi_t^{(i)} = \psi^{(i)}(\epsilon_t)$, $1 \leq i \leq v$. Наложим на них условия

$$\varphi \in (0, 1), \quad D\psi^{(i)}(\xi_t) = d(y) < \infty, \quad y \in Y. \quad (1)$$

Пусть $I^{(i)}$ — интервалы на числовой прямой и Y_I — множество всех тех и только тех управлений, для которых

$$W_{\varphi^{(i)}}(y) = E\psi^{(i)} \in I^{(i)} = (a_i, b_i), \quad i = 1, \dots, v.$$

Цель управления: максимизировать на множестве Y_I средний выигрыш $W_\varphi(y) = E\varphi$. Более точную формулировку цели дадим для случая, когда пространство управлений $Y = \{y_1, \dots, y_k\}$ конечно: для любой пары положительных чисел ϵ и δ построить такую управляющую систему $A_{\epsilon, \delta}$, что

$$W(A_{\epsilon, \delta}; \varphi) > \max_{y \in Y_I} E\varphi - \epsilon,$$

где $W(A_{\epsilon, \delta}; \varphi)$ — предельное математическое ожидание функционала φ по мере, порожденной стратегией $A_{\epsilon, \delta}$, и,

$$\sum_{i \in Y_I} \pi_i < \delta \sum_{j \in Y_I} \pi_j,$$

где π_i — предельная вероятность совершения действия y_i , вычисленная по той же мере, $Y_I = Y - Y_I$.

Иными словами, система $A_{\epsilon, \delta}$ должна преимущественно избирать действия из Y , и при этом максимизировать $E\varphi$. Допустим, что эта цель достижима.

Решаем эту задачу в «адаптивной постановке». Рассматривается класс ОПНЗ с одинаковыми пространствами X и Y и функционалами $\varphi, \psi^{(1)}, \dots, \psi^{(v)}$, удовлетворяющими (1), и требуется построить ϵ -оптимальное семейство адаптивных систем. Их конструкция основывается на тех же принципах, что и автоматы $\tilde{D}_{k,n}$; они обозначаются $(DC)_{k,n}$.

Перефразируя высказанные выше условия на объект управления, введем класс $B_{k,v}$ процессов. Ими являются векторные (($v+1$)-мерные) ОПНЗ $\eta_t = (\varphi_t, \psi_t^{(1)}, \dots, \psi_t^{(v)})$, компоненты которых подчинены условиям (1). Тем самым нет необходимости требовать наблюдаемости ξ , и управлению подлежат наблюдаемые векторные ОПНЗ класса $B_{k,v}$.

Входными сигналами $(DC)_{k,n}$ служат наборы вида $(\varphi(\xi_t), \psi^{(1)}(\xi_t), \dots, \psi^{(v)}(\xi_t))$, а выходными — элементы Y . Множество состояний состоит из центрального состояния s_0 и k групп периферических состояний S_1, \dots, S_k . В s_0 равновероятно выбирается $y_i \in Y$, которое повторяется $m=m(n)$ тактов подряд. По реализации процесса вычисляются оценки математических ожиданий

$$\sum_{\psi^{(j)}, m} (y_i) = \frac{\psi^{(j)}(\xi_1) + \dots + \psi^{(j)}(\xi_m)}{m}$$

и проверяются включения $\Sigma_{\psi^{(j)}, m} (y_i) \in I_{\epsilon}^{(i)} = (a_i + \epsilon, b_i - \epsilon)$, $i=1, \dots, v$, где $\epsilon = \epsilon(n)$. Если хоть одно из включений нарушено, сохраняется состояние s_0 и процедура повторяется. Если же все включения верны, то автомат переходит в группу состояний $S_i = \{s_l^i, l=1, \dots, n\}$, которым сопоставлено действие y_i . Переходы в каждой группе S_i происходят так же, как и в ветвях автомата $\tilde{D}_{k,n}$. Во всяком состоянии s_l^i действие y_i избирается m раз и подсчитывается эмпирический средний выигрыш

$$\sum_{\varphi, m} (y_i) = \frac{\varphi(\xi_1) + \dots + \varphi(\xi_m)}{m},$$

который принадлежит интервалу $(0, 1)$. Эта величина преобразуется в поощрение или в наказание, первое с вероятностью $\Sigma_{\varphi, m} (y_i)$, а второе с дополнительной вероятностью. Поощрение переводит s_l^i в s_n^i , а наказание s_l^i в s_{l-1}^i (из s_l^i в центральное состояние s_0).

Относительно последовательностей $m(n)$ и $\epsilon(n)$ предполагается следующее:

$$\lim_{n \rightarrow \infty} m(n) = \infty, \quad \lim_{n \rightarrow \infty} \epsilon(n) = 0.$$

Нам потребуется более полная характеристика функции $m(n)$, которую запишем в виде $m(n) = \alpha(n) \beta(n)$. Здесь оба сомножителя неограниченно растут вместе с n и подчиняются ограничениям

$$\alpha(n) \varepsilon^2(n) \leq c < \infty, \quad \lim_{n \rightarrow \infty} \frac{2^{an}}{\beta(n)} = 0, \quad a > 0. \quad (2)$$

Второе условие означает, что $\beta(n)$ можно принять, например, в виде $\beta(n) = 2^{n^2}$.

При управлении ОПНЗ ξ_t с помощью автомата $(DC)_{k,n}$ этот процесс оказывается марковским. Допустим, что он эргодический. Это выполняется, в частности, когда X принадлежит числовой прямой и существует плотность вероятностей $p(x|y)$, ограниченная снизу $p(x|y) \geq p_0 > 0$ всюду на X при любом $y \in Y$. Эргодический процесс имеет предельное распределение, не зависящее от начального состояния. Поэтому существует предельный средний выигрыш $W((DC)_{k,n}; \varphi)$ автоматов $(DC)_{k,n}$, т. е. предел математических ожиданий E_φ .

Теорема 2. Для процессов класса $B_{k,n}$, автоматы $(DC)_{k,n}$ являются ε -оптимальным семейством адаптивных систем, т. е. для любых положительных ε , δ существует такое $n_{\varepsilon,\delta}$, что при всех $n > n_{\varepsilon,\delta}$ автоматы $(DC)_{k,n}$ обеспечивают цель управления типа условного экстремума.

Доказательство. Воспользуемся известной нам (§ 2 гл. III) связью средних времен совершения действий с предельными вероятностями действий

$$\frac{T^i(n)}{T^j(n)} = \frac{\pi_i}{\pi_j},$$

которая сохраняет силу для $(DC)_{k,n}$. Средние времена вычисляются с помощью системы разностных уравнений

$$T_z^i = p_i T_n^i + (1 - p_i) T_{z-1}^i + m, \quad z = 1, \dots, n,$$

$$T_0^i = m + T_1^i q_i$$

и равны $T^{(i)}(n) = T_0^i$. В этих уравнениях T_z^i — среднее время до смены действия, если автомат находится в состоянии s_z^i . Коэффициенты в уравнении означают: p_i — вероятность увеличения глубины состояния автомата при действии y_i , q_i — вероятность перехода из центрального состояния s_0 в состояние s_1^i (для простоты положим $v=1$). Они соответственно равны

$$p_i = E_\varphi(y_i), \quad q_i = P(\Sigma_{\varphi,m}(y_i) \in I).$$

Решая систему разностных уравнений, находим

$$T_n^i = m(n) \left[1 + \frac{1}{(1 - p_i)^n} P(\Sigma_{\varphi,m}(y_i) \in I) \right].$$

Если $y_i \in Y_I$, то $\lim_{n \rightarrow \infty} P(\Sigma_{\psi, m(n)} \in I) = 1$; а если $y_i \notin Y_I$, то по неравенству Чебышева находим

$$(1 - p_i)^{-n} P(\Sigma_{\psi, m(n)} \in I) \leq \frac{d}{m(n)(1 - p_i)^n \varepsilon^2(n)},$$

где $d = \max_y d(y)$. Используя вид $m(n)$ и $\varepsilon(n)$ и условия (2), наклоненные на них, получаем при $y_i \notin Y_I$, $y_j \in Y_I$

$$\lim_{n \rightarrow \infty} \frac{T^{(i)}(n)}{T^{(j)}(n)} = 0.$$

В случае, когда y_i и y_j принадлежат Y_I , но $p_i < p_j$, имеем такое же равенство.

Из этих равенств получаем

$$\lim_{n \rightarrow \infty} W((DC_k, n); \varphi) = \max_{y \in Y_i} E_\varphi(y).$$

Если n_δ выбрано так, что при всех $n > n_\delta$

$$W((DC)_k, n; \varphi) \geq \max_{y \in Y_I} E_\varphi(y) - \varepsilon,$$

а целое n_δ таково, что при всех $n > n_\delta$ выполняется неравенство

$$\sum_{i \in \bar{Y}_I} T^{(i)}(n) < \delta \sum_{j \in Y_I} T^{(j)}(n),$$

то зададим $n_\varepsilon, \delta = \max(n_\delta, n_\delta)$. Тогда при всяком $n > n_\varepsilon, \delta$ автоматы $(DC)_k, n$ обеспечивают цель управления. Теорема доказана.

ГЛАВА V

РЕКУРРЕНТНЫЕ ПРОЦЕДУРЫ УПРАВЛЕНИЯ ОПНЗ

§ 1. Постановка задачи

В этой главе преимущественно рассматривается класс \mathfrak{R} ОПНЗ, у которых пространствами, фазовым X и управлений Y , служит числовая прямая. Предполагается, что при всех y существует и конечен средний выигрыш $W(y) = \int_{-\infty}^{\infty} x \mu(dx | y)$. С помощью этой функции высказана цель управления. Процессы из \mathfrak{R} наблюдаемые.

Мы хотим построить адаптивные системы, обеспечивающие в классе \mathfrak{R} назначенную цель. Имея в виду простоту управляющих систем для ОПНЗ в случаях, когда известны свойства процесса, следует адаптивные системы также искать в достаточно простом виде. Можно быть уверенными, что стратегии общего вида $\sigma = \{F_t(\cdot | x^t, y^{t-1}), t \geq 1\}$ для управления ОПНЗ не требуются. Стратегии упрощаются, если они нерандомизированные и основываются на прошлом фиксированной глубины, т. е. задаются правилами вида $f_t(x_{t-l}^t, y_{t-l}^{t-1})$, которые не требуют слишком большой памяти для хранения предыстории процесса. Отметим, что часто эти рекуррентные процедуры имеют глубину $1 - y_t = f_t(x_t, y_{t-1})$. Но и при больших глубинах реализация таких процедур на ЦВМ не встречает затруднений.

В случае общей рекуррентной процедуры вида $y_t = f_t(x_{t-l}, \dots, x_t; y_{t-l}, \dots, y_{t-1})$ следует еще задавать начальные значения y_0, y_1, \dots, y_{l-1} . Далее рассматриваются стратегии, порожденные такими процедурами. Это означает, что двумерный процесс (ξ_t, y_t) записывается следующим образом: $(\xi_t(y_{t-1}), y_t(\xi_{t-l}^t, y_{t-l-1}^{t-1}))$, т. е. представляет собой l -связный марковский процесс. Способ вычисления действий должен быть таким, что возникающее блуждание по пространству Y приводит к достиже-

нию цели. Если сформулированная цель достигается стационарной программной стратегией, повторением «оптимального» действия y_0 , то можно перефразировать цель в терминах блуждания по Y : выполнить предельное соотношение $\lim_{t \rightarrow \infty} y_t = y_0$, понимаемое в надлежащем смысле (например, сходимость в среднем квадратическом).

Допустим, что искомая адаптивная система существует. Она представима обучаемой системой вида $L = (X, S, Y; T)$, где X и Y — множества входных и выходных сигналов, S — множество состояний, образованное компонентами, $S = (D, R_l, X^l \times Y^l)$, которые означают следующее: D — множество правил, R_l — пространство значений l -мерной статистики $\zeta_t = (\xi_t, \xi_{t-1}, \dots, \xi_{t-l})$ и $X^l \times Y^l$ — «память» системы, содержит предысторию глубины l — набор $(\xi_t, y_{t-1}, \xi_{t-1}, y_{t-2}, \dots, \xi_{t-l}, y_{t-l})$. Наконец, $T = T_{\zeta_t, t} = f_t$ — семейство преобразований правил выбора действий или, что в нашем случае означает то же самое, преобразования самих действий.

Подчеркнем, что управляющая система, представляющая собой рекуррентную процедуру, реализует обычно нестационарную стратегию.

Принятую рекуррентную процедуру вычисления действий $y_t = f_t(\xi_{t-l}^t, y_{t-l-1}^{t-1})$ будем именовать «методом стохастической аппроксимации» или, кратко, МСА.

§ 2. Задача о выполнении плана

Пусть R_m — класс всех ОПНЗ с пространствами, фазовым X и управлений Y , — числовыми прямыми и конечными средними выигрышами $W(y)$ — монотонными функциями. Пусть существует и единственный корень y^* уравнения $W(y) = W^*$, где W^* — фиксированное число, интерпретируемое в практических задачах как плановое задание.

Зададимся целью управления — достигнуть заданный уровень W^* среднего выигрыша. Переформулируем ее в терминах случайного блуждания по пространству Y : построить такую процедуру вычисления действий y_t по

наблюдениям за ОПНЗ из класса R_M , чтобы с вероятностью 1 $\lim_{t \rightarrow \infty} y_t = y^*$.

В допущении (для определенности), что $W(y)$ — монотонно убывающая функция, дадим «вывод» процедуры вычисления искомой последовательности действий y_t , который одновременно служит доказательством сходимости на физическом уровне строгости. В основе рассуждения лежит бесспорный факт, что среднее значение величины $\xi_{t+1}(y_t) - W^*$ равно $W(y_t) - W^*$. Эта разность положительна слева от (неизвестной) точки y^* и отрицательна справа от нее. Если предыдущие шаги рассматриваемой процедуры привели нас, отправляясь от начального значения y_0 , к действию y_t , то естественно принять за следующее действие y_{t+1} сумму $y_t + a(\xi_{t+1}(y_t) - W^*)$, где $a > 0$. Будем считать коэффициент a меняющимся со временем, т. е. полагаем $a = a(t)$. Таким образом, приходим к рекуррентным соотношениям

$$y_{t+1} = y_t + a(t)[\xi_{t+1}(y_t) - W^*], \quad t \geq 0, \quad (1)$$

где значение y_0 фиксировано (иногда его удобно принять случайным, но так, чтобы $Ey_0^2 < \infty$). Начальная точка y_0 может находиться сколь угодно далеко от искомого y^* , следовательно, частичные суммы ряда $\sum_{t=1}^{\infty} a(t)$ должны безгранично расти, обеспечивая возможность приблизиться от y_0 к y^* . Потребуем еще, чтобы $\lim_{t \rightarrow \infty} a(t) = 0$. Практически несомненно, что если y_0 расположено от y^* слева, то величина $\xi_{t+1}(y_t) - W^*$ положительна по крайней мере на первых шагах использования равенства (1) и y_t окажется приближающейся к y^* . В дальнейшем среди этих разностей появляются как положительные, так и отрицательные. Поэтому члены последовательности y_t иногда удаляются от y^* , но в силу положительности в среднем величины $\xi_{t+1}(y_t) - W^*$ они имеют тенденцию стремиться к y^* . В некоторый момент выполнится неравенство $y_t > y^*$, и тогда разность $\xi_{t+1}(y_t) - W^*$ в среднем окажется отрицательной. Равенство (1) указывает, что y_{t+1} уменьшится по сравнению с y_t , т. е. члены нашей последовательности начнут приближаться к y^* убывая. В силу того, что

$a(t) \rightarrow 0$, скачки y_t вокруг y^* могут затухать, хотя из-за существования разброса $\xi_{t+1}(y_t)$ относительно среднего значения разности $\xi_{t+1}(y_t) - W^*$ могут быть сколь угодно большими по абсолютной величине. Нужную нам сходимость $y_t \rightarrow y^*$ обеспечивает условие $\sum_{t=1}^{\infty} a^2(t)$, приводящее к подавлению «помех».

Таким образом, процедура (1) требует знания поведения функции $W(y)$ — факта ее убывания. В случае возрастаания $W(y)$ следует пользоваться такой процедурой:

$$y_{t+1} = y_t - a(t)[\xi_{t+1}(y_t) - W^*], \quad t \geq 0. \quad (2)$$

Без дополнительных оговорок далее предполагаем выполнеными условия

$$a(t) > 0, \quad \sum_{t=1}^{\infty} a(t) = \infty, \quad \sum_{t=1}^{\infty} a^2(t) < \infty. \quad (3)$$

Легко понять, что порождаемые процедурами (1) и (2) (процедурами Роббинса — Монро, кратко ПРМ) случайные последовательности y_t представляют собой марковские процессы. Это непосредственно вытекает, например, из марковости последовательности $\xi_{t+1}(y_t)$, ибо y_t получается из нее линейным преобразованием.

Высказанные выше наводящие соображения о сходимости ПРМ могут быть подкреплены строгим доказательством при выполнении надлежащих допущений о классе ОПНЗ.

Обозначим через $\mathcal{R}' \subset R_M$ класс ОПНЗ, удовлетворяющих условиям:

1) существует число $d > 0$ такое, что при всех y

$$\mathbb{E}\xi^2(y) \leq d(1 + y^2);$$

2) при любом $\epsilon > 0$ справедливо либо неравенство

$$\sup_{\epsilon < |y-y^*| < \epsilon^{-1}} (W(y) - W^*)(y - y^*) < 0,$$

либо

$$\inf_{\epsilon < |y-y^*| < \epsilon^{-1}} (W(y) - W^*)(y - y^*) > 0.$$

Заметим, что из первого условия следует, что средний выигрыш не может расти быстрее, чем линейная функция $|W(y)| \leq d_1(1 + |y|)$.

Мы считаем, что известно, удовлетворяет ОПНЗ первому или второму неравенству условия 2). В первом случае пользуемся вариантом (1) ПРМ, а во втором — (2).

Теорема 1. Для всех ОПНЗ из класса \mathfrak{R}' при любом начальном значении y_0 процедура Роббинса—Монро обеспечивает достижение цели, т. е.

$$\mathbf{P} \left(\lim_{t \rightarrow \infty} y_t = y^* \right) = 1.$$

Доказательство. Примем для определенности, что $W(y)$ не убывает в окрестности корня и $y^* = W^* = 0$. Это означает, что $y_{t+1} = y_t - a(t) \xi_{t+1}(y_t)$. Отсюда из условия 1) имеем

$$\begin{aligned} \mathbf{E}(y_{t+1}^2 | y^t) &= y_t^2 - 2a(t)y_t W(y_t) + a^2(t) \mathbf{E}(\xi_t^2 | y^t) \leq \\ &\leq y_t^2 - 2a(t)y_t W(y_t) + a^2(t)d(1 + y_t^2). \end{aligned} \quad (4)$$

Согласно условию 2) $\mathbf{E}(y_{t+1}^2 | y^t) \leq y_t^2(1 + da^2(t)) + da_2(t)$. Введем новые случайные величины

$$z_t = y_t^2 \prod_{j=t}^{\infty} (1 + da^2(j)) + d \sum_{j=t}^{\infty} a^2(j) \prod_{m=j+1}^{\infty} (1 + da^2(m)). \quad (5)$$

Для условных математических ожиданий легко проверяется неравенство $\mathbf{E}(z_{t+1} | z^t) \leq z_t$, из которого следует

$$\dots \leq \mathbf{E}z_t \leq \dots \leq \mathbf{E}z_1 < \infty. \quad (6)$$

Отсюда заключаем, что z_t — полумартингал, значит, $\mathbf{P} \left(\lim_{t \rightarrow \infty} z_t = z_{\infty} \right) = 1$ и, следовательно, $\mathbf{P} \left(y_t^2 \xrightarrow{t \rightarrow \infty} \bar{y} \right) = 1$.

Теорема будет доказана, если мы покажем, что $\bar{y} = 0$ с вероятностью 1. Для этого из предыдущего сначала выводим, что ограничены $\mathbf{E}y_t^2$ (см. (5) и (6)). Возьмем математические ожидания от обеих частей неравенства (4) для $t = 1, 2, \dots$ и, складывая первые t из них, приходим к неравенству

$$\mathbf{E}y_{t+1}^2 \leq \mathbf{E}y_0^2 + d \sum_{j=1}^t a^2(j)[1 + \mathbf{E}y_j^2] - 2 \sum_{j=1}^t a(j) \mathbf{E}(y_j W(y_j)).$$

Ряд $\sum_{j=1}^{\infty} a_j \mathbf{E}(y_j W(y_j))$ должен сходиться (иначе правая часть начиная с некоторого t окажется отрицательной), а это в силу $\sum_{j=1}^{\infty} a^2(j) < \infty$ и условия 2) означает, что найдется подпоследовательность y_t , для которой $\mathbf{P}(y_{t_j} W(y_{t_j}) \rightarrow 0) = 1$. Сопоставляя это с условием 2) и сходимостью $y_t^2 \rightarrow \tilde{y}$, приходим к выводу, что $\mathbf{P}(\tilde{y} = 0) = 1$.

Эта теорема означает, что для любого $\epsilon > 0$ найдется такой случайный момент времени τ_ϵ , конечный с вероятностью 1, что при всех $t > \tau_\epsilon$ имеем

$$|y_t - y^*| < \epsilon.$$

Было бы полезно знать вероятностные характеристики этого немарковского момента, такие, как его распределение вероятности, моменты $\mathbf{E}\tau_\epsilon^m$, а также асимптотика хотя бы первых двух моментов при стремящемся к нулю ϵ .

В практически важном случае непрерывной функции $W(y)$ сходимость последовательности y_t влечет за собой сходимость соответствующих средних выигрышей $\lim_{t \rightarrow \infty} W(y_t) = W^*$ (в том же смысле, что и y_t). Справедливо более сильное утверждение: с вероятностью 1

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_1^t \xi_j = W^*.$$

Итак, марковский процесс y_t , полученный из ОПНЗ ξ_t с помощью ПРМ, сходится к значению y^* . Какова скорость сходимости? Будем судить о ней по убыванию среднего квадрата разности $y^* - y_t$. Примем простоты ради функцию $W(y)$ монотонно убывающей и воспользуемся такой конкретной формой ПРМ:

$$y_{t+1} = y_t + \frac{a}{t} [\xi_{t+1}(y_t) - W_0^*], \quad t \geq 1. \quad (7)$$

где число a положительно.

Теорема 2. Если выполнены условия:

- $W(y) - W^* \leq -\lambda(y - y^*)$ при всех y ,
- $E\xi_t^2(y) \leq d(1 + y^2)$, то $E(y^* - y_t)^2 \sim \frac{c}{t}$, $c > 0$.

Более полные сведения о предельных свойствах y_t доказываются благодаря дополнительным ограничениям на ОПНЗ ξ_t . Условимся обозначать через $\eta_t(y) = \xi_t(y) - W(y)$ центрированный процесс ($E\eta_t \equiv 0$), отвечающую ему меру обозначаем P .

Теорема 3. Пусть ОПНЗ ξ_t с монотонно убывающей функцией $W(y)$ удовлетворяет следующим условиям:

- $E\xi^2(y) \leq d(1 + y^2)$ для всех y ;
- $W(y) = W^* - \alpha(y - y^*) + o(|y - y^*|)$ при $y \rightarrow y^*$, причем $\alpha > 1/(2a)$;
- существует и конечен предел $\lim_{y \rightarrow y^*} E\eta_t^2(y) = \sigma^2$;
- для некоторого $v > 0$

$$\lim_{H \rightarrow \infty} \sup_{|y-y^*| < v} \int_{|\eta| > H} \eta_t^2(y) dP = 0.$$

Тогда при любом фиксированном начальном значении y^* последовательность y_t , порожденная ПРМ (7), асимптотически нормальна, ее математическое ожидание y^* , а дисперсия

$$\frac{a^2\sigma^2}{(2a\alpha - 1)t}, \text{ m. e. } \sqrt{t}(y_t - y^*) \sim N\left(0, \frac{a^2\sigma^2}{2a\alpha - 1}\right).$$

Указываемая теоремами 2 и 3 скорость сходимости порядка $t^{-1/2}$ характерна для процессов математической статистики и не может быть повышена.

В предыдущей постановке задачи и теоремах предполагалось, что оптимальное действие y^* единственное. Возникает вопрос о поведении последовательности y_t в противном случае, т. е. когда уравнение $W(y) = W_0^*$ имеет несколько различных корней. Можно ли рассчитывать на попадание y_t в множество $B = \{y : W(y) = W_0^*\}$? Допустим, что для функции $W(y)$ множество B состоит из конечного числа связных компонент. Используем обо-

значение $U_{\epsilon, 1/\epsilon}(B) = V_\epsilon(B) \cap \left\{y : |y| < \frac{1}{\epsilon}\right\}$, где $V_\epsilon(B)$ — ϵ -окрестность множества B .

Теорема 4. *Пусть существует неотрицательная функция $u(y)$ с непрерывной и ограниченной второй производной $u''(y)$, причем $\lim_{|y| \rightarrow \infty} u(y) = 0$ такая, что*

$$\sup_{y \in U_{\epsilon, 1/\epsilon}(B)} (W(y) - W_0) u'(y) < 0, \quad \epsilon > 0,$$

$$E\xi^2(y) \leq d(1 + u(y)).$$

Тогда с вероятностью 1 y_t сходится либо к одной из точек B , либо к границе одной из его связных компонент.

В случае нескольких корней уравнений $W(y) = W^*$ предельное множество последовательности y_t можно несколько сузить по сравнению с приведенной теоремой, если наложить на ОПНЗ ряд ограничений.

Изложенные выше результаты о сходимости и асимптотических свойствах ПРМ допускают распространение на многомерный случай. Это означает, что рассматривается векторный (l -мерный) ОПНЗ $\xi_t = (\xi_t^{(1)}, \dots, \xi_t^{(l)})$, управлением которого в каждый момент времени являются l -мерные векторы $y = (y^{(1)}, \dots, y^{(l)})$. Допустим, что все компоненты имеют математические ожидания $W^{(i)}(y) = E\xi_t^{(i)}(y)$, $i = 1, \dots, l$, образующие вектор $W(y) = (W^{(1)}(y), \dots, W^{(l)}(y))$. Задача о выполнении плана состоит в отыскании такого y^* , что $W(y^*) = W^*$ — заданный постоянный вектор. Ее решение методом стохастической аппроксимации дает непосредственное обобщение ПРМ

$$y_{t+1} = y_t + a(t)(\xi_{t+1}(y_t) - W^*), \quad t \geq 0,$$

начальное значение y_0 фиксировано (либо случайно, но тогда $E|y_0|^2 < \infty$). В предположениях, которые естественно обобщают формулировки приведенных выше теорем, доказываются соответствующие свойства последовательности y_t , а именно, сходимость с вероятностью 1, оценка скорости сходимости, асимптотическая нормальность и особенности в случае нескольких корней.

§ 3. Задача о максимизации выигрыша

Пусть \mathfrak{R}_d — класс ОПНЗ, у которых фазовое пространство X и пространство управлений Y совпадают с чистовой прямой и существует непрерывно дифференцируемый средний выигрыш $W(y)$ с единственным максимумом. Абсциссу максимума обозначим y_{opt} , т. е. $W(y_{\text{opt}}) = \max_y W(y)$, и требуем, чтобы при $y < y_{\text{opt}}$ $W(y)$ строго возрастила, $W'(y) > 0$, а при $y > y_{\text{opt}}$ строго убывала, $W'(y) < 0$.

Целью управления выдвинем максимизацию среднего выигрыша. В терминах блуждания по Y цель перефразируется так: указать алгоритм построения последовательности y_t , который для всякого управляемого процесса приводит с вероятностью 1 к равенству $\lim_{t \rightarrow \infty} y_t = y_{\text{opt}}$. Ясно,

что это влечет асимптотическую оптимальность в слабом смысле $P(\lim_{t \rightarrow \infty} W(y_t) = W(y_{\text{opt}})) = 1$. Нетрудно убедиться,

что достигается также асимптотическая оптимальность в сильном смысле $P\left(\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{l=1}^t \xi_l(y_{l-1}) = W(y_{\text{opt}})\right) = 1$.

Обратимся к построению алгоритма, проводя рассуждения на интуитивном уровне.

В высказанных предположениях задача максимизации $W(y)$ равносильна задаче решения уравнения $W'(y) = 0$. Поэтому воспользуемся идеей рекуррентной процедуры Роббинса—Монро увеличения или уменьшения текущего значения действия y в зависимости от того, положительна или нет оценка производной $W'(y)$. Статистическую оценку $W'(y)$ строим из наблюдаемых значений управляемого процесса ξ_t :

$$\frac{\xi(y + c) - \xi(y - c)}{2c}$$

при достаточно малых $c > 0$. Математическое ожидание этого отношения с точностью до исчезающе малого (вместе

с c) слагаемого равно $W'(y)$. Значит, процедура пересчета действий записывается в форме

$$\tilde{y} = y + a \frac{\xi(y+c) - \xi(y-c)}{2c}, \quad (1)$$

смысл которой очевиден. Остается распорядиться свободными параметрами a и c так, чтобы была нужная сходимость действий. Считаем их зависящими от t , $a = a(t)$, $c = c(t)$ и подчиненными условиям:

a) $a(t) > 0$, $c(t) > 0$, $\lim_{t \rightarrow \infty} c(t) = 0$;

б) $\sum_{t=1}^{\infty} a(t) = \infty$;

в) $\sum_{t=1}^{\infty} a(t)c(t) < \infty$;

г) $\sum_{t=1}^{\infty} \frac{a^2(t)}{c^2(t)} < \infty$.

Их роль обсудим позднее, а сейчас заметим лишь, что из этих условий следует $\lim_{t \rightarrow \infty} a(t) = 0$. Удобно выбирать эти последовательности в таком виде:

$$a(t) = a/t, \quad a > 0, \quad c(t) = \frac{e}{t^\gamma}, \quad e > 0, \quad 0 < \gamma < 1/2. \quad (3)$$

Правило преобразований действий во времени получим из (1), приписав следующий порядок наблюдений за процессом ξ_t и выполнений управляемых действий

$$y_{2t+2} = y_{2t} + \frac{a(t)}{c(t)} [\xi_{2t+2}(y_{2t} + c(t)) - \xi_{2t+1}(y_{2t} - c(t))]. \quad (4)$$

где y_0 фиксировано. Процедуру вычисления y_t согласно (4) назовем *процедурой Кифера — Вольфовича* (кратко ПКВ). Обратим внимание на то, что действиями в моменты $2t$ и $2t+1$ служат не y_{2t} и y_{2t+1} , а $y_{2t} - c(t)$ и $y_{2t} + c(t)$ соответственно. Величина y пересчитывается лишь

в четные моменты времени и является абсциссой точки, в которой оценивается значение производной $W'(y)$.

Благодаря условию б) в (2) от начального y_0 возможно приблизиться к любому y_{opt} . С ростом t , когда $c(t) \rightarrow 0$, отношение $\frac{\xi(y_{2t} + c(t)) - \xi(y_{2t} - c(t))}{2c(t)}$ в среднем все более точно изображает производную $W'(y_{2t})$, которая стремится в свою очередь к 0, если $y_{2t} \rightarrow y_{opt}$. Условия в) и г) в (2) означают, что $c(t)$ должны убывать ни слишком быстро, ни слишком медленно.

Последовательность y_t представляет собой 2-связный марковский процесс. Сформулируем условия его сходимости к значению y_{opt} .

Введем класс $\mathfrak{R} \subset \mathfrak{R}_d$ ОПНЗ, удовлетворяющих требованиям:

1) существует число $d > 0$ такое, что при всех y

$$(W'(y))^2 + E\xi^2(y) \leq d(1 + y^2);$$

2) непрерывная на числовой прямой функция $W'(y)$ удовлетворяет условию Липшица.

Теорема 1. Для всех ОПНЗ из класса \mathfrak{R} процедура Кифера — Вольфовича при всяком начальном y_0 обеспечивает достижение цели, т. е.

$$P\left(\lim_{t \rightarrow \infty} y_t = y_{opt}\right) = 1.$$

Доказательство теоремы опускаем.

Отсюда следует, что существует такая, почти наверное, конечная, случайная величина τ_ϵ , что при всех $t > t_\epsilon$

$$W(y_t) > W(y_{opt}) - \epsilon.$$

Было бы полезно знать свойства (хотя бы математическое ожидание) этого немарковского момента, указывающие скорость приближения y_t к y_{opt} .

Воспользуемся среднеквадратическим критерием оценки скорости сходимости ПКВ, т. е. функцией $v(t) = E(y_t - y_{opt})^2$. Оценка ее зависит от дифференциальных свойств $W(y)$ в окрестности y_{opt} . Если считать $W(y)$

всего лишь имеющей непрерывную производную, то при задании $a(t)$ и $c(t)$ в виде (3) оказывается

$$v(t) = \begin{cases} o\left(\frac{1}{t^{1-2\gamma}}\right), & \gamma \geq \frac{1}{4}, \\ o\left(\frac{1}{t^{2\gamma}}\right), & \gamma < \frac{1}{4}. \end{cases}$$

Следовательно, полученная оценка дает лучший результат при $\gamma = 1/4$. Более того, при $\gamma \neq 1/4$ существуют такие меры $\mu(\cdot | y)$, для которых $\lim_{t \rightarrow \infty} v(t) t^{1/2-\epsilon} > 0$ при некотором $\epsilon > 0$. Если $W''(y)$ непрерывна (в окрестности y_{opt}), то

$$v(t) = \begin{cases} o\left(\frac{1}{t^{1-2\gamma}}\right), & \gamma \geq \frac{1}{6}, \\ o\left(\frac{1}{t^{4\gamma}}\right), & \gamma < \frac{1}{6}, \end{cases}$$

и порядок убывания $v(t)$ можно сделать не хуже $o(t^{-2/3})$. Наконец, если в окрестности максимума функция $W(y)$ аналитична и симметрична, то

$$v(t) = o(t^{-(1-2\gamma)})$$

при всех $\gamma \in (0, 1/2)$.

Таким образом, ПКВ оказывается медленно сходящейся процедурой. Дополним приведенные оценки скорости следующим фактом относительно асимптотической нормальности последовательности действий y_t . При различных (и притом достаточно общих) предположениях о виде, свойствах и гладкости функции $W(y)$ последовательность

$$t^{1/2-\gamma} (y_t - y_{\text{opt}})$$

или

$$t^{1/2} c(t) (y_t - y_{\text{opt}})$$

асимптотически нормальна с нулевым математическим ожиданием и дисперсией, которая явным образом выражается через некоторые числовые характеристики управляемого ОПНЗ.

В заключение распространим ПКВ на ОПНЗ, у которых пространством управлений Y служит m -мерное пространство, т. е. $\mathbf{y} = (y^{(1)}, \dots, y^{(m)})$, с компонентами — вещественными числами. Задача о максимизации $W(\mathbf{y})$ решается посредством такой рекуррентной процедуры

$$\mathbf{y}_{t+2m} = \mathbf{y}_t + \frac{a(t)}{c(t)} \Delta_c \xi_t,$$

где компоненты вектора приращений $\Delta_c \xi_t = (\Delta \xi_t^{(1)}, \dots, \Delta \xi_t^{(m)})$ определены формулами

$$\Delta \xi_t^{(i)} = \xi_{t+2i} (\mathbf{y}_t + c(t) \mathbf{e}_i) - \xi_{t+2i-1} (\mathbf{y}_t - c(t) \mathbf{e}_i), \\ i = \overline{1, m},$$

в которых $\mathbf{e}_i = (\delta_{ij}, j = 1, \dots, m)$ — единичные векторы. Очевидно, эта процедура есть обобщение (4). С увеличением размерности Y приходится больше времени тратить на «изучение» свойств функции $W(\mathbf{y})$. На многомерную ПКВ распространяются сформулированные выше результаты.

§ 4. Приложение рекуррентных процедур к задаче прогноза стационарных последовательностей

Лежащие в основе стохастической аппроксимации идеи часто и с успехом используются в разнообразных приложениях. Отправной точкой подобных приложений является редукция рассматриваемой проблемы к решению уравнения или отысканию максимума. Обоснование сходимости полученной процедуры, если оно не сводится просто к упоминанию теорем из §§ 2, 3 (или их аналогов), встречает обычно серьезные трудности и тогда об эффективности алгоритма судят на основании разумности и естественности получаемых с его помощью результатов.

Мы рассмотрим здесь такую проблему: наблюдается траектория стационарного процесса ξ_t , и в каждый момент t известны ν значений этого процесса $\xi_{t-\nu+1}, \dots, \xi_{t-1}, \xi_t$. Требуется оценить будущее значение процесса $\xi_{t+\tau}$, по возможности с меньшей ошибкой ($t_0 \geq 1$). Иными словами, мы ищем такой алгоритм прогноза на t_0 шагов вперед по предыстории глубины ν , который в пределе (по времени) приводил бы к наименьшей возможной среднеквадратической ошибке в заданном классе процессов. Примем в качестве такого класса гауссовские процессы с $E \xi_t = 0$ и рациональной спектральной плотностью, а оценку будущего значения ξ_{t+1} изберем в виде линейной формы

$$\xi_{t+1} = y^{(1)} \xi_t + y^{(2)} \xi_{t-1} + \dots + y^{(\nu)} \xi_{t-\nu+1}, \quad (1)$$

где $y^{(1)}, \dots, y^{(v)}$ — коэффициенты, выбираемые из условия минимальности функции

$$E(\xi_t - \hat{\xi})^2 = D(y^{(1)}, \dots, y^{(v)}) = E(\xi_{t+1} - y^{(1)}\xi_t - \dots - y^{(v)}\xi_{t-v+1})^2.$$

Если корреляционная функция процесса $R(m)$ известна, «оптимальный» прогнозатор находится без труда: надо решить систему линейных уравнений

$$\frac{\partial D}{\partial y^{(i)}} = 0, \quad i = 1, \dots, v. \quad (2)$$

В аддитивной постановке задачи прогноза корреляционная функция неизвестна и можно располагать лишь величинами ошибок прогноза. Этим обстоятельством мы воспользуемся для уточнения коэффициентов линейной формы (1), минимизирующих квадратичную форму $D(y^{(1)}, \dots, y^{(v)})$, являющихся корнями системы (2). Согласно ПРМ вектор коэффициентов $y_t = (y_t^{(1)}, \dots, y_t^{(v)})$ находится посредством рекуррентных соотношений

$$y_{t+1} = y_t - 2^{-1}a(t) \nabla_y (\xi_{t+1} - y_t^{(1)}\xi_t - \dots - y_t^{(v)}\xi_{t-v+1})^2$$

или

$$y_{t+1}^{(i)} = y_t^{(i)} + a(t) (\xi_{t+1} - y_t^{(1)}\xi_t - \dots - y_t^{(v)}\xi_{t-v+1}) \xi_{t-i}, \quad i = 1, \dots, v,$$

т. е. прогнозирующая система нелинейная. Функционал, являющийся оценкой будущего значения процесса, имеет третью степень относительно наблюдаемых значений процесса.

В частном случае прогноза на единицу времени по предыстории глубины 1 имеем

$$y_{t+1} = y_t + \frac{a}{t} (\xi_{t+1} - y_t \xi_t) \xi_t, \quad a > 0.$$

Аналитическое исследование аддитивной прогнозирующей системы, основывающейся на этом методе вычисления коэффициента $y^{(1)}$, вызывает затруднения, и поэтому мы ограничимся результатами

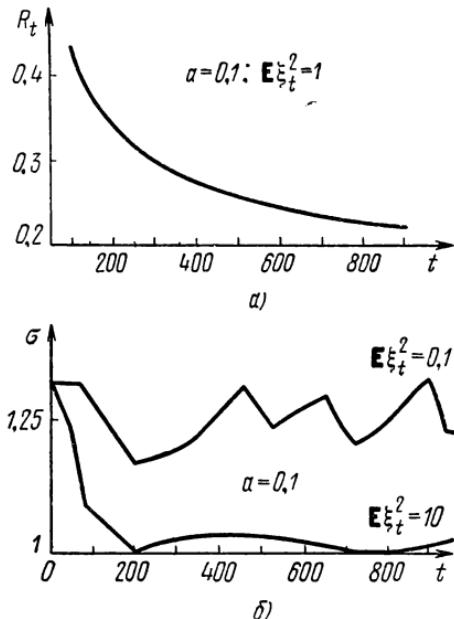


Рис. 9.

моделирования процесса прогнозирования на ЦВМ. В проведенных численных экспериментах с многими стационарными процессами избранного класса подсчитывались средние значения величин $(y_t - y_{\text{opt}})^2$ и $(\xi_t - \hat{\xi}_t)^2$ по 100 реализациям через каждые 100 тактов на временном отрезке длины от 1000 и более.

На рис. 9, а и 9, б представлена динамика величин *

$$R_t = \frac{(y_t - y_{\text{opt}})^2}{(y_1 - y_{\text{opt}})^2} \quad \text{и} \quad \sigma(t) = \left(\frac{\overline{(\xi_t - \hat{\xi}_t)^2}}{E(\xi_t - \hat{\xi}_t, \text{opt})^2} \right)^{1/2}$$

в зависимости от связи чисел a и $d = E \xi_t^2$. Мы видим, что при $ad=1$ имеет место убывание $\overline{(y_t - y_{\text{opt}})^2}$, по-видимому, степенного характера $\overline{(y_t - y_{\text{opt}})^2} \sim (y_1 - y_{\text{opt}})^2 t^{-\alpha}$ со значениями a , близкими к единице. Относительно последовательности значений $(\xi_t - \hat{\xi}_t)^2$ можно утверждать, что при том же условии ($ad=1$) она сходится к минимуму среднеквадратической ошибки. Эти данные указывают на то, что вместо коэффициента a/t целесообразно положить $a(t) = \left(\sum_1^t \xi_i^2 \right)^{-1}$, что автоматически учитывает дисперсию прогнозируемого стационарного процесса.

§ 5. Стохастическое программирование

Через B обозначим класс векторных ОПНЗ $\xi_t = (\xi_t^{(1)}, \dots, \xi_t^{(v)})$ с вещественными компонентами. Пространство действий Y также векторное, его элементы $y = (y^{(1)}, \dots, y^{(v)})$. Допустим, что все процессы из B имеют вектор средних выигрышей

$$W(y) = \int_{-\infty}^{\infty} x \mu(dx | y),$$

относительно которого сформулирована задача типа условного экстремума. Ниже на одном примере указывается, как распространить методы стохастической аппроксимации на такие задачи. Изложение относится к скалярному процессу ξ_t с пространством действий $Y = R$, — r -мерным евклидовым пространством.

Требуется найти максимум среднего выигрыша $W(y)$ в выпуклом замкнутом множестве $M \subset Y$. Относительно

*). Чертка означает образование эмпирической оценки.

$W(\mathbf{y})$ предполагается непрерывность, ограниченность сверху и выпуклость. Для достижения этой цели воспользуемся «обобщенным градиентом» функции $W(\mathbf{y})$, которым является любой вектор $\hat{W}(\mathbf{y})$, удовлетворяющий неравенству

$$W(\mathbf{y}) - W(\mathbf{z}) \leq (\hat{W}(\mathbf{y}), \mathbf{y} - \mathbf{z})$$

при любых \mathbf{y} и \mathbf{z} .

Обозначим через ζ_t векторный r -мерный процесс, который при всех t удовлетворяет условию

$$\mathbf{E}(\zeta_t | \mathbf{y}^t) = \alpha_t \hat{W}(\mathbf{y}_t) + \beta_t,$$

где α_t — случайные величины, β_t — случайные векторы. Такой процесс ζ_t называют *стохастическим квазиградиентом* в точке \mathbf{y}_t . Введем ограниченную случайную величину η_t , удовлетворяющую неравенствам

$$\mathbf{E}(\|\zeta_t\|^2 | \mathbf{y}^t) \leq \gamma^2 \eta_t^2 \leq k_c$$

при $\|\mathbf{y}_l\| \leq c$, $l = 1, \dots, t$, где число $c < \infty$ любое.

Последовательность управлений определена рекуррентным образом:

$$\mathbf{y}_{t+1} = \pi_M(\mathbf{y}_t - a(t) b(t) \zeta_t), \quad t \geq 1, \quad (1)$$

где $\pi_M(z)$ — оператор проектирования точки z на множество M , ввиду высказанного условия на M определен однозначно, $a(t)$ и $b(t)$ — последовательности неотрицательных чисел. Наложим следующие ограничения:

$$a(t) \geq 0, \quad \alpha_t \geq 0, \quad \sum_{t=1}^{\infty} a(t) \mathbf{E} \|\beta_t\| < \infty,$$

$$\sum_{t=1}^{\infty} a^2(t) < \infty, \quad 0 < \underline{b} \leq b(t)(\eta_t + \lambda_t \|\mathbf{y}_t\|) \leq \bar{b} < \infty,$$

$$\mathbf{P}\left(\sum_{t=1}^{\infty} a(t) \alpha_t = \infty\right) = 1,$$

где \underline{b} и \bar{b} — числа,

$$\lambda_t = \begin{cases} 1, & \|\beta_t\| > 0, \\ 0, & \beta_t = 0. \end{cases}$$

Справедлив следующий результат.

Т е о р е м а 1. *При высказанных предположениях*

$$\mathsf{P} \left(\lim_{t \rightarrow \infty} \mathbf{y}_t = \mathbf{y}_{\text{opt}} \right) = 1,$$

где \mathbf{y}_{opt} — решение (вообще говоря, не единственное) рассматриваемой задачи.

Изложенная здесь схема допускает распространение на следующую задачу. Максимизировать средний выигрыш одной компоненты векторного ОПНЗ:

$$W^{(1)}(\mathbf{y}) = \mathsf{E}^{\xi^{(1)}}$$

при соблюдении ограничений на остальные:

$$W^{(j)}(\mathbf{y}) = \mathsf{E}^{\xi^{(j)}} \geqslant 0, \quad j = 2, \dots, v.$$

Относительно надлежащим образом составленной рекуррентной процедуры доказывается, что она сходится с вероятностью 1 к решению этой задачи.

ГЛАВА VI

АВТОМАТНЫЕ МЕТОДЫ ДЛЯ ВЕКТОРНЫХ ОПНЗ

§ 1. Децентрализованное управление. Теоретико-игровая интерпретация

Пусть на траекториях ОПНЗ ξ_t с пространствами, фазовым X и управлений Y , определены v функционалов $\varphi_j(\xi_t)$, $1 \leq j \leq v$, и при стратегиях заданного класса существуют математические ожидания $W^{(i)} = E_{\varphi_i}$. Цели управления ОПНЗ сформулированы в терминах набора $(W^{(1)}, \dots, W^{(v)})$. Требуется синтезировать адаптивную систему, обеспечивающую достижение «многокритериальной» цели. Такие системы уже встречались нам в векторной задаче о выполнении плана и задачах типа условного экстремума (в главах III и V). В этой главе рассматривается обратная проблема — проблема анализа, — заключающаяся в определении целей, которые обеспечиваются обучаемыми системами определенной структуры. Исследуемые системы являются адаптивными для этих целей. В случаях, когда за обучаемые системы приняты конечные автоматы, средствами разрешения поставленной проблемы анализа оказываются изучение предельного поведения ассоциированной марковской цепи и моделирование на ЦВМ взаимодействия автоматов с управляемым процессом.

Исходным объектом управления служит ОПНЗ ξ_t в фазовом пространстве X , а цель относится к функционалам $\varphi_t^{(j)} = \varphi^{(j)}(\xi_t)$. Удобно и естественно заменить этот процесс на v -мерный векторный $\Phi_t = (\varphi_t^{(1)}, \dots, \varphi_t^{(v)})$, который также есть ОПНЗ. Сменив обозначения на привычные, — вместо φ_t вводим вектор $\xi_t = (\xi_t^{(1)}, \dots, \xi_t^{(v)})$ — рассматриваем теперь задачу управления векторным процессом $\xi_t = (\xi_t^{(1)}, \dots, \xi_t^{(v)})$ с фазовым пространством X , принадлежащим v -мерному евклидову пространству R_v . Еще несколько усложним объект. Будем считать пространство управлений векторным, $Y = Y_1 \times Y_2 \times \dots \times Y_r$, и

управление в момент t изображать вектором $\mathbf{y}_t = (y_t^{(1)}, \dots, y_t^{(v)})$. Эволюция процесса ξ_t регулируется распределением

$$\mu(M | \mathbf{y}_{t-1}) = \mu(\xi_t^{(1)} \in M_1, \xi_t^{(2)} \in M_2, \dots, \xi_t^{(v)} \in M_v | \mathbf{y}_{t-1}),$$

$$M = M_1 \times M_2 \times \dots \times M_v.$$

Математическое ожидание функционала $\psi(\xi_t)$ равно

$$W(\mathbf{y}) = \int_{\mathbf{x}} \psi(x) \mu(dx | \mathbf{y}),$$

т. е. является функцией r аргументов. Мы в качестве функционалов на ξ_t выбираем компоненты вектора $(\xi_t^{(1)}, \dots, \xi_t^{(v)})$. Их математические ожидания обозначим

$$W^{(j)}(\mathbf{y}) = \int_{\mathbf{x}} x_j \mu(dx_1, \dots, dx_j, \dots, dx_v | \mathbf{y}).$$

Цель управления должна формулироваться в терминах этих функций.

Зададим вид системы управления описанным классом процессов.

Элементарными обучаемыми системами назовем объекты

$$L_i = (Z_i, S_i, Y_i; T_{\zeta}^{(i)}, t), \quad i = 1, \dots, r,$$

где $Z_i = ([\xi]^i)$ — подмножество компонент процесса ξ_t , поступающих на вход системы, S_i — множество состояний, включающих в себя правила выбора действий, память и значения статистики ζ , $T_{\zeta}^{(i)}$ — семейство операторов на множестве правил.

Из определения следует, что система L_i вырабатывает i -ю компоненту вектора \mathbf{y} . Мы примем, что объединение номеров компонент процесса, поступающих на входы всех r систем, равно всем номерам $1, 2, \dots, v$, т. е. каждая компонента процесса поступает на вход хотя бы одной системы.

Прямым произведением обучаемых систем L_i , $i = \overline{1, r}$, называется обучаемая система

$$L = L_1 \times \dots \times L_r = (X, S, Y; T_{\zeta}, t),$$

где X — множество значений v -мерных векторов ξ_t (управляемого процесса), $S = S_1 \times \dots \times S_r$, $Y = Y_1 \times \dots \times Y_r$, $T_{\zeta, t}$ — оператор на S , который зависит от статистики $\zeta = (\zeta^{(1)}, \dots, \zeta^{(r)})$ и осуществляет отображение $T_{\zeta, t}(s^{(1)}, \dots, s^{(r)}) = (T_{\zeta^{(1)}, t}s^{(1)}, \dots, T_{\zeta^{(r)}, t}s^{(r)}), s^{(i)} \in S_i$.

Под децентрализованной системой управления процессом ξ_t с векторным пространством управлений $Y = Y_1 \times \dots \times Y_r$ понимают прямое произведение обучаемых систем L_1, \dots, L_r . Управление такого типа означает параллельное действие обучаемых систем, каждая из которых ответственна за свою компоненту вектора $y = (y^{(1)}, \dots, y^{(r)})$.

Пусть K — класс ОПНЗ, компоненты которого принимают значения из конечного множества, Y конечно. Тогда в качестве обучаемых систем естественно рассматривать автоматы.

Предметом настоящей главы является исследование децентрализованных систем управления векторными ОПНЗ и установление достигаемых целей управления при различных предположениях о классе ОПНЗ и строении автоматов — элементарных обучаемых систем. Оказывается удобным употребление терминологии теории игр.

Партией называется вектор $y = (y^{(1)}, \dots, y^{(r)}), y^{(i)} \in Y_i$. *Исходом* партии y называется значение ξ_t ОПНЗ, реализовавшееся после выбора $y_{t-1} = y$. Вероятности исходов партии в случае конечного пространства X равны вероятностям $\mu(\xi_t | y)$.

Игра состоит из неограниченной последовательности партий и их исходов. При $v > 2$ говорят об играх многих лиц. Обучаемые системы L_1, \dots, L_r называются *участниками игры*. В важнейшем случае, когда эти системы суть автоматы, мы говорим об играх автоматов. Эта антропоморфная терминология не должна приводить к недоразумениям.

Вектор *платежных функций* определен равенством

$$W(y) = \int_X x \mu(dx | y),$$

его j -й компонентой является введенное ранее математическое ожидание j -й компоненты ОПНЗ при условии, что

приложено управление y . Нас главным образом будет интересовать случай конечного фазового пространства X . Тогда интегралы в выражении для $W(y)$ заменяются суммами. Эти последние упрощаются, если ОПНЗ имеет независимые компоненты, т. е. условные вероятности имеют вид

$$\mu((\xi^{(1)}, \dots, \xi^{(v)}) | y) = \prod_{j=1}^v \mu_j(\xi^{(j)} | y).$$

Для таких процессов платежные функции равны

$$W^{(j)}(y) = \sum_{x \in X} x \mu_j(x | y), \quad j = 1, \dots, v.$$

Перечислим несколько характерных типов партий.

Ситуацией равновесия (партией Нэша) назовем партию $y_R = (y_R^{(1)}, \dots, y_R^{(v)})$ такую, что $W^{(j)}(y_R) \geq W^{(j)}(y_R^{(1)}, \dots, y_R^{(j-1)}, y, y_R^{(j+1)}, \dots, y_R^{(v)})$, $y \neq y_R^{(j)}$ при всех j . Иными словами, смена действия j -м участником игры влечет за собой его потери.

Решением называется такая партия, в которой все платежные функции достигают максимума. Ясно, что решение есть ситуация равновесия.

Скажем, что j -й участник игры *критический* для партии y , если $W^{(j)}(y) = \min_i W^{(i)}(y)$. Партия y_M называется *максминной*, если

$$\min_i W^{(i)}(y_M) = \max_y \min_i W^{(i)}(y).$$

Иными словами, партия — максминная, если отвечающий ей критический участник игры в среднем «получает» не меньше, чем критический участник любой другой партии.

Понятия ситуации равновесия и максминной партии, вообще говоря, не включаются одно в другое. Ясно, что ситуация равновесия не обязана быть максминной партией и, наоборот, максминная партия — не обязательно ситуация равновесия. В отдельных случаях, однако, эти партии тесно связаны. Укажем один пример.

В игре «в размещения» участники имеют по одинаковому числу действий y_1, \dots, y_k и задается она упорядоченными положительными числами a_1, \dots, a_k , $1 \geq a_1 \geq \dots \geq a_2 \geq \dots \geq a_k > 0$. Пусть \mathbf{y}' — партия, в которой v_j

участников избрали действие y_j ($v_j \geq 0, \sum_{j=1}^k v_j = v$), тогда

каждый из них получает доход 1 («поощрение») с вероятностью a_j/v_j и доход 0 («наказание») с дополнительной вероятностью. Исход партии \mathbf{y} определяется набором чисел (v_j) . Игра в размещения имеет ситуации равновесия, каждой из которых отвечают партии с набором (v_j) , удовлетворяющим неравенствам при любых парах $i, j = \overline{1, k}$:

$$\frac{a_i}{v_i} \geq \frac{a_j}{v_j + 1}.$$

Покажем, что каждая ситуация равновесия этой игры является максиминной партией.

Пусть \mathbf{y}_R — ситуация равновесия. Никакому участнику игры, в том числе критическому, невыгодно изменить действие, т. е.

$$\min_l W^{(l)}(\mathbf{y}_R) \geq \frac{a_j}{v_j(\mathbf{y}_R) + 1}, \quad j = 1, \dots, k.$$

Рассмотрим произвольную партию \mathbf{y} такую, что $v_j(\mathbf{y}) \neq v_j(\mathbf{y}_R)$ хотя бы при одном j . Найдется хоть одно j_0 такое, что $v_{j_0}(\mathbf{y}) \geq v_{j_0}(\mathbf{y}_R) + 1$. Отсюда следует

$$\min_l W^{(l)}(\mathbf{y}) \leq \frac{a_{j_0}}{v_{j_0}(\mathbf{y})} \leq \frac{a_{j_0}}{v_{j_0}(\mathbf{y}_R) + 1} \leq \min_l W^{(l)}(\mathbf{y}_R).$$

Значит, $\min_l W^{(l)}(\mathbf{y}_R) = \max_y \min_l W^{(l)}(\mathbf{y})$, т. е. \mathbf{y}_R является максиминной партией.

Легко привести примеры игр в размещения, у которых имеются максиминные партии, не являющиеся ситуациями равновесия.

Ниже всюду предполагается, что размерности управляемого векторного ОПНЗ и вектора управления одинаковы:

$$v = r.$$

Кроме того, считаем компоненты процесса независимыми.

§ 2. Игры бинарных автоматов.

Аналитические методы

Рассматривается класс v -мерных ОПНЗ с независимыми бинарными компонентами и пространством управлений $Y = Y^{(1)} \times Y^{(2)} \times \dots \times Y^{(v)}$, причем все $Y^{(j)} = (y_1^{(j)}, \dots, \dots, y_{k_j}^{(j)})$ конечные. Каждый такой процесс $\xi_t = (\xi_t^{(1)}, \dots, \xi_t^{(v)})$ задают совокупностью $\{q_j(y), j = 1, \dots, v, y \in Y\}$ вероятностей

$$q_j(y) = P(\xi_t^{(j)} = 1 | y)$$

значения 1 (поощрения) j -й компоненты в партии y , тогда числа

$$p_j(y) = 1 - q_j(y)$$

равны вероятностям значения 0, которое будем теперь сопоставлять наказанию. Математические ожидания компонент (т. е. средние выигрыши или «платежные функции») равны $W^{(j)}(y) = q_j(y)$.

В качестве обучаемой системы изберем прямое производение v конечных автоматов из ϵ -оптимальных семейств бинарных автоматов

$$A^{(n)} = A_1^{(n)} \times \dots \times A_v^{(n)},$$

причем все автоматы — «сомножители» — имеют одинаковую глубину памяти n . Каждый из этих автоматов представляет собой набор $A_i^{(n)} = \{X, S_i^{(n)}, Y_i; \Pi^{(i, n)}\}$, где $X = \{0, 1\}$, $S_i^{(n)}$ — множество состояний (из $k_i n$ элементов), а $\Pi^{(i, n)} = (\Pi_{+}^{(i, n)}, \Pi_{-}^{(i, n)})$ — две стохастические матрицы вероятностей переходов между состояниями: $\Pi_{+}^{(i, n)} = \|p_{ij}^{(i)}(+)\|$ —

матрица переходов при поощрениях (+), $\Pi_{-}^{(l, n)} = \| p_{i,j}^{(l)}(-) \|$ — при наказаниях (—). Все автоматы мурковские.

Сопоставим объекту $A^{(n)} \otimes \xi$ ассоциированную марковскую цепь $M_n = (S^{(n)}, \mathcal{P}^{(n)})$. Ее множество состояний $S^{(n)} = S_1^{(n)} \times \dots \times S_v^{(n)}$, а элементы матрицы $\mathcal{P}^{(n)}$, которые обозначим $p_{\bar{\lambda}, \bar{\mu}}$ (где $\bar{\lambda}$ и $\bar{\mu}$ — индексы состояний $s_{\bar{\lambda}} = (s_{\lambda_1}, \dots, s_{\lambda_v})$ и $s_{\bar{\mu}} = (s_{\mu_1}, \dots, s_{\mu_v})$ из $S^{(n)}$), вычисляются по формуле

$$p_{\bar{\lambda}, \bar{\mu}} = \prod_{i=1}^v [q_i(\mathbf{y}) p_{\lambda_i, \mu_i}(+) + p_i(\mathbf{y}) p_{\lambda_i, \mu_i}(-)],$$

где $\mathbf{y} = (y^{(1)}, \dots, y^{(v)})$ — действие автомата $A^{(n)}$, отвечающее его состоянию $s_{\bar{\lambda}}$, т. е. $y^{(j)}$ — выходной сигнал $A_j^{(n)}$, находящегося в состоянии s_{λ_j} . Легко видеть, что $\mathcal{P}^{(n)}$ — стохастическая матрица.

Нас будут интересовать случаи, когда цепь M_n эргодическая. Если в игре принимают участие автоматы из последовательностей $D_{k_1, n}, K_{k_1, n}, L_{k_1, n}$ или иные из числа рассмотренных в §§ 2, 3 гл. III, эргодичность имеет место при условии, что для каждого ОПНЗ выполнено условие:

при любых $i = 1, \dots, v$ и $\mathbf{y} \in Y$

$$0 < W^{(i)}(\mathbf{y}) < 1.$$

Суждения о свойствах автоматов $A^{(n)}$ как обучаемых систем по отношению к классу бинарных ОПНЗ, о достижимых целях управления мы сделаем после отыскания предельных вероятностей партий (таким же образом, как в § 2 гл. III). Однако надо иметь в виду, что вычисление и даже оценка этих вероятностей — трудоемкая задача. Действительно, если автоматы $A_1^{(n)}, \dots, A_v^{(n)}$ имеют соответственно по k_1, \dots, k_v действий, каждому из которых отвечает цепочка из n состояний, то цепь M_n содержит $|S^{(n)}| = k_1 \dots k_v n^v$ состояний.

Пусть каким-нибудь способом найдены предельные вероятности $\pi^{M_n}(s)$ состояний в цепи M_n , тогда предельные вероятности партий получаются суммированием:

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

где $S^{(n)}(\mathbf{y}) \subset S^{(n)}$ — множество всех тех и только тех состояний, которым отвечает партия \mathbf{y} . Множество $S^{(n)}$ допускает представление в виде объединения непересекающихся подмножеств

$$S^{(n)} = \bigcup_{\mathbf{y} \in Y} S^{(n)}(\mathbf{y}).$$

Введем еще одно понятие.

Предельным средним выигрышем j -го автомата — участника игры — называется

$$W_n^{(j)} = \sum_{\mathbf{y} \in Y} W^{(j)}(\mathbf{y}) \pi_n(\mathbf{y}).$$

Нас будут интересовать асимптотические свойства вектора $(W_n^{(1)}, \dots, W_n^{(v)})$ при неограниченном росте памяти n . Так, если окажется, что при надлежащих условиях для некоторой партии \mathbf{y}_0 имеем $\lim_{n \rightarrow \infty} \pi_n(\mathbf{y}_0) = 1$, можно заключить, что при всех j оказывается $W_n^{(j)} \xrightarrow[n \rightarrow \infty]{} W^{(j)}(\mathbf{y}_0)$. Таким образом, следует изучить последовательность ассоциированных марковских цепей M_n , $n = -1, 2, \dots$. Возникающие аналитические трудности можно будет преодолеть, если построить цепи более простые, чем M_n , но с эквивалентными в подходящем смысле асимптотическими свойствами. Такое упрощение возможно осуществить, пользуясь следующим методом.

Пусть M — марковская цепь с множеством состояний S . Выделим подмножество $S' \subset S$ и будем рассматривать переходы между состояниями в моменты попадания в S' . Они образуют новую марковскую цепь M' — «сужение M на множество S' ». Один такт функционирования цепи M' соответствует нескольким тактам исходной цепи M между двумя попаданиями в состояния подмножества S' . Можно доказать, что если цепь M эргодическая, то M' также эргодическая и предельные вероятности состояний из S' этих цепей связаны равенством

$$\pi^{M'}(s) = \frac{\pi^M(s)}{\sum_{s_i \in S'} \pi^M(s_i)}, \quad s \in S'. \quad (1)$$

Отсюда, в частности, вытекает, что для любой пары состояний $s_i, s_j \in S'$ имеем

$$\frac{\pi^M(s_i)}{\pi^M(s_j)} = \frac{\pi^{M'}(s_i)}{\pi^{M'}(s_j)}.$$

Возвратимся к ассоциированным марковским цепям $M_n = (S^{(n)}, \mathcal{P}^{(n)})$. В каждом из $x = k_1 k_2 \dots k_n$, подмножеств $S^{(n)}(y)$, на которые разлагается $S^{(n)}$, выделим по одному состоянию — тому, которому отвечают состояния максимальной глубины n всех автоматов. Эти состояния — «глубокие» — обозначим s_1, \dots, s_x ; они образуют множество Γ . Введем марковскую цепь M'_n : сужение цепи M_n на множество Γ всех глубоких состояний. Отметим, что все эти цепи имеют одно и то же фиксированное множество состояний, различаясь лишь значениями матриц вероятностей переходов. Предельные вероятности состояний цепи M_n обозначаются $\pi^{M_n}(s)$.

Дальнейший анализ относится к играм автоматов, в которых участвуют автоматы $D_{k_1, n}, D_{k_2, n}, \dots, D_{k_n, n}$ с одинаковой памятью n и разными, вообще говоря, числами действий. Это ограничение сделано для того, чтобы упростить формулировки результатов и их доказательства. Итак, всюду в этом параграфе слово автомат означает автомат $D_{k, n}$.

Теорема 1. *Существуют положительные числа c_1 и c_2 такие, что для $j=1, \dots, n$ и всех $n \geq 1$ справедливы неравенства*

$$c_1 \leq \frac{\pi_n(y_j)}{\pi^{M'_n}(s_j)} \leq c_2.$$

Доказательство. Прежде всего установим, что для каждого $j=1, \dots, n$ найдется такое $d_j > 1$, что при всех n

$$\pi_n(y_j) < d_j \pi^{M_n}(s_j), \quad (2)$$

т. е. отношение предельной вероятности самого «глубокого» состояния s_j из множества $S^{(n)}(y_j)$ к предельной вероят-

ности всего $S^{(n)}(\mathbf{y}_j)$ не стремится к нулю при неограниченном росте n . Это неравенство удобно записать в виде

$$\sum_{s \in S^{(n)}(\mathbf{y}_j)} \frac{\pi^{M_n}(s)}{\pi^{M_n}(\mathbf{s}_j)} < d_j$$

и для доказательства (2) нужно сначала оценить $\frac{\pi^{M_n}(s)}{\pi^{M_n}(\mathbf{s}_j)}$.

С этой целью зафиксируем n и состояние $s \in S^{(n)}(\mathbf{y}_j)$. Через G_{nj} обозначим подмножество всех граничных состояний $S^{(n)}(\mathbf{y}_j)$, т. е. состояний, из которых за один шаг можно покинуть $S^{(n)}(\mathbf{y}_j)$. Введем марковскую цепь F_n — сужение цепи M_n на множество $\mathbf{s}_j \cup s \cup G_{nj}$. Справедливы соотношения

$$\frac{\pi^{M_n}(s)}{\pi^{M_n}(\mathbf{s}_j)} = \frac{\pi^{F_n}(s)}{\pi^{F_n}(\mathbf{s}_j)} \leqslant \frac{p_{\mathbf{s}_j s}^{F_n} + \sum_{s_i \in G_{nj}} p_{\mathbf{s}_j s_i}^{F_n}}{p_{s s_j}^{F_n}}, \quad (3)$$

в которых правое неравенство вытекает из неравенства

$$\frac{\pi(s_i)}{\pi(s_j)} \leqslant \frac{1 - p_{jj}}{p_{ij}}.$$

Оценим вероятности, фигурирующие в правой части (3). Если все автоматы одновременно получили поощрение, в цепи M_n совершается переход из любого $s \in S^{(n)}(\mathbf{y}_j)$ в глубокое состояние \mathbf{s}_j . Отсюда следует

$$p_{s s_j}^{F_n} \geqslant \prod_{i=1}^v q_i. \quad (4)$$

Если состоянию s цепи M_n отвечают состояния глубины l_1, \dots, l_v автоматов $A_1^{(n)}, \dots, A_v^{(n)}$, то при некотором $\gamma > 0$

$$p_{\mathbf{s}_j s}^{F_n} \leqslant \gamma \prod_{i=1}^v p_i^{n-l_i}. \quad (5)$$

Для установления этого неравенства введем ряд обозначений: $r_{\mathbf{s}_j s}(t)$ — вероятность перехода в цепи M_n из \mathbf{s}_j в s

ровно за t шагов (без захода в этом промежутке в s_j и в $G_{n,j}$), $\Delta = \max_i (n - l_i)$. Ясно, что менее чем за Δ шагов нельзя перейти из s_j в s , поэтому

$$p_{s_j s}^{F_n} = \sum_{t=\Delta}^{\infty} r_{s_j s}(t).$$

В цепи M_n рассмотрим множество траекторий $s_0^t = (s(0), s(1), \dots, s(t))$, начинающихся каждая в состоянии $s_j = s(0)$ и оканчивающихся спустя t шагов в $s = s(t)$, причем $s(1), \dots, s(t-1)$ отличны от s_j и не принадлежат $G_{n,j}$. Сопоставим всякой такой траектории траекторию $[s_0^t]$, вообще говоря, более короткую: она до момента $t^* = t - \Delta$ совпадает с s_0^t , а в момент $t^* + 1$ оказывается в исходном состоянии s_j . Легко видеть, что отношение вероятности любой траектории $[s_0^t]$ с началом и концом в состоянии s_j к суммарной вероятности всех траекторий, которым она поставлена в соответствие, равно $\prod_{i=1}^v q_i / p_i^{n-l_i}$ и, следовательно, имеем

$$\frac{r_{s_j s}(t)}{r_{s_j s_j}(t^*)} = \prod_{i=1}^v \frac{p_i^{n-l_i}}{q_i}.$$

Отсюда непосредственно вытекает равенство

$$p_{s_j s}^{F_n} = \sum_{t=\Delta}^{\infty} r_{s_j s}(t) = \prod_{i=1}^v \frac{p_i^{n-l_i}}{q_i} \sum_{t^*=1}^{\infty} r_{s_j s_j}(t^*) = p_{s_j s_j}^{F_n} \prod_{i=1}^v \frac{p_i^{n-l_i}}{q_i},$$

которое влечет за собой справедливость неравенства (5).

Пусть s — граничное состояние, ему отвечает состояние глубины 1 хотя бы одного автомата. Для таких состояний из (5) следует оценка

$$p_{s_j s}^{F_n} \leq \gamma (\max_i p_i)^{n-1}, \quad s \in G_{n,j}.$$

В правую часть неравенства (3) подставим эту оценку, а также (4) и (5). Суммируя полученные неравенства по

всем $s \in S^{(n)}(\mathbf{y}_j)$, приходим к (2). Сверху того, очевидно неравенство

$$\pi^{\mathbf{M}_n}(s_j) \leq \pi_n(\mathbf{y}_j). \quad (6)$$

Введем обозначения: $c_2 = \max_j d_j > 1$, $c_1 = 1/c_2$. Суммируя обе части неравенства (2), находим, что

$$\sum_{j=1}^x \pi^{\mathbf{M}_n}(s_j) > c_1.$$

Это дает, в силу (1), при всех j

$$c_1 \pi^{\mathbf{M}'_n}(s_j) < \pi^{\mathbf{M}_n}(s_j) < \pi^{\mathbf{M}'_n}(s_j).$$

Отсюда согласно (2) и (6) получаем утверждение теоремы.

Согласно теореме 1 предельные вероятности состояний s_j цепи M'_n лишь постоянными множителями отличаются от предельных вероятностей партий \mathbf{y}_j . Поэтому возможно заменить сложные марковские цепи M_n с безгранично растущими количествами состояний на существенно более простые цепи M'_n с одним и тем же фиксированным числом состояний x . Оценим вероятности переходов $p_{ij}^{M'_n}$ в цепи M'_n .

Теорема 2. Существуют положительные числа c_3 , c_4 такие, что если партии \mathbf{y}_i и \mathbf{y}_j отличаются действиями автоматов с индексами m_1, \dots, m_r , то при всех n

$$c_3 \prod_{l=1}^r [1 - W^{(m_l)}(\mathbf{y}_j)]^n \leq p_{ij}^{M'_n} \leq c_4 [\max_i (1 - W^{(l)}(\mathbf{y}_i))]^n.$$

Доказательство. Через H_n обозначим марковскую цепь — сужение цепи M_n на множество $\Gamma \cup G_{n,i}$. Поскольку для перехода из s_i в s_j цепь M_n непременно должна оказаться во множестве граничных состояний $G_{n,i}$, имеем $p_{ij}^{M'_n} \leq \sum_{s \in G_{n,i}} p_{si}^{H_n}$. Пользуясь оценкой (5) и суммируя, получаем правую часть доказываемого неравенства.

Переходя к доказательству левой части неравенства, примем исключительно для упрощения записи, что партии \mathbf{y}_i , \mathbf{y}_j различаются действиями лишь двух автоматов A_{m_1} и A_{m_2} . Общий случай рассматривается аналогично.

Отождествим действие $y_j^{(m)}$ автомата A_m с целым числом j , и пусть в этом случае $y_j^{(m_1)} - y_i^{(m_1)} = d_1 \pmod{k_{m_1}}$ и $y_j^{(m_2)} - y_i^{(m_2)} = d_2 \pmod{k_{m_2}}$.

Переход в цепи M_n' из одного глубокого состояния в другое за $n+d_1+d_2$ шагов (минуя остальные глубокие) может происходить многими способами. Один из вариантов состоит в следующем: автомат A_{m_1} получит подряд $n+d_1-1$ наказаний, а затем d_2+1 поощрений, автомат A_{m_2} получит d_1 поощрений, $n+d_2-1$ наказаний и затем одно поощрение. Пусть все остальные автоматы за это время не смешили свои действия. Для этого достаточно, чтобы в моменты времени, кратные n , они получили поощрения. Теперь можно указать оценку снизу вероятности $p_{\bar{s}_i \bar{s}_j}^{M_n'}$:

$$p_{\bar{s}_i \bar{s}_j}^{M_n'} \geq \left[p_{m_1}^{d_1-1} q_{m_1}^{d_2+1} p_{m_2}^{d_2-1} q_{m_2}^{d_1} \prod_{i \neq m_1, m_2} q_i \right] (p_{m_1}, p_{m_2})^n,$$

из которой находим окончательный результат. Теорема полностью доказана.

Общий метод применения полученных результатов состоит в попарном сравнении предельных вероятностей состояний цепи M_n' . При этом полезно использовать уже знакомое неравенство

$$\frac{\pi_{\bar{s}_j \bar{s}_j}^{M_n'}(\bar{s}_i)}{\pi_{\bar{s}_i \bar{s}_j}^{M_n'}(\bar{s}_j)} \leq \frac{1 - p_{\bar{s}_j \bar{s}_j}^{M_n'}}{p_{\bar{s}_i \bar{s}_j}^{M_n'}}$$

и получаемую с помощью теоремы 2 оценку

$$1 - p_{\bar{s}_j \bar{s}_j}^{M_n'} = \sum_{i \neq j} p_{\bar{s}_j \bar{s}_i}^{M_n'} \leq (x-1) c_4 \left[\max_i [1 - W^{(1)}(\bar{y}_i)] \right]^n. \quad (7)$$

Далее везде в примерах партии нумеруются лексикографически.

Пример 1. В игре принимают участие два автомата $D_{2,n}$. Платежные функции принимают такие значения:

$$W^{(1)}(\bar{y}_1) = 0,6, \quad W^{(1)}(\bar{y}_2) = 0,75, \quad W^{(1)}(\bar{y}_3) = 0,4,$$

$$W^{(1)}(\bar{y}_4) = 0,9, \quad W^{(2)}(\bar{y}_1) = 0,7, \quad W^{(2)}(\bar{y}_2) = 0,55,$$

$$W^{(2)}(\bar{y}_3) = 0,3, \quad W^{(2)}(\bar{y}_4) = 0,5.$$

Отношения предельных вероятностей состояний цепи M'_n оцениваются следующим образом:

$$\frac{\pi_{M'_n}(s_2)}{\pi_{M'_n}(s_1)} \leq \left(\frac{0,4}{0,45}\right)^n, \quad \frac{\pi_{M'_n}(s_3)}{\pi_{M'_n}(s_1)} \leq \left(\frac{0,4}{0,6}\right)^n,$$

$$\frac{\pi_{M'_n}(s_4)}{\pi_{M'_n}(s_1)} \leq \frac{\pi_{M'_n}(s_4)}{\pi_{M'_n}(s_3)} \cdot \frac{\pi_{M'_n}(s_3)}{\pi_{M'_n}(s_1)} \leq \left(\frac{0,7}{0,5} \cdot \frac{0,4}{0,6}\right)^n.$$

Согласно теореме 1 отсюда находим $\lim_{n \rightarrow \infty} \pi_n(y_r) = 0$, $r = 2, 3, 4$. Следовательно, $\lim_{n \rightarrow \infty} \pi_n(y_1) = 1$. Таким образом, при достаточно большой памяти один автомат в среднем «получает» выигрыш, близкий к 0,6, а другой — к 0,7.

Поставим в соответствие игре автоматов Γ ориентированных граф V_Γ : число его вершин равно числу партий в игре, из вершины i в вершину j проходит дуга, если партии y_i и y_j отличаются действием ровно одного автомата и этот автомат критический для партии y_i .

Лемма. *Если на графике игры вершина j достижима из вершины i , то при некотором $c > 0$ для всех n*

$$\frac{\pi_n(y_i)}{\pi_n(y_j)} \leq c \left[\frac{\max_i (1 - W^{(i)}(y_j))}{\max_i (1 - W^{(i)}(y_i))} \right]^n.$$

Доказательство. Определим последовательность марковских цепей M''_n , $n = 1, 2, \dots$, с тем же множеством состояний, что у цепей M'_n , и с переходными вероятностями

$$P_{j_1 j_2}^{M''_n} = \begin{cases} \frac{P_{j_1 j_2}^{M'_n}}{1 - P_{j_1 j_1}^{M'_n}}, & \text{если } j_1 \neq j_2, \\ 0, & \text{если } j_1 = j_2. \end{cases} \quad (7')$$

Предельные вероятности состояний цепей M'_n и M''_n связаны равенством

$$\frac{\pi_{M'_n}(s_i)}{\pi_{M'_n}(s_j)} = \frac{\pi_{M''_n}(s_i)}{\pi_{M''_n}(s_j)} \cdot \frac{1 - P_{j_j}^{M'_n}}{1 - P_{i_i}^{M'_n}}. \quad (8)$$

Если на графе игры V_G существует дуга из вершины i_1 в вершину i_2 , то из определения графа и теоремы 2 следуют оценки

$$c_3 \left[\max_l [1 - W^{(l)}(\mathbf{y}_{i_1})] \right]^n \leq p_{i_1 i_2}^{M'_n} \leq c_4 \left[\max_l (1 - W^{(l)}(\mathbf{y}_{i_1})) \right]^n,$$

правая из которых, вместе с неравенством (7), дают нам $p_{i_1 i_2}^{M''_n} \geq \frac{c_3}{(\kappa - 1) c_4} = c > 0$, а значит, $\pi^{M''_n}(s_{i_2}) \geq c \pi^{M''_n}(s_{i_1})$.

В предположении достижимости j -й вершины из i -й, найдется связывающий их путь длины, не превышающей κ . Значит,

$$\pi^{M''_n}(s_j) \geq c \pi^{M''_n}(s_i). \quad (9)$$

Из равенства $1 - p_{ii}^{M'_n} = \sum_{j \neq i} p_{ij}^{M'_n}$, того факта, что на графике из i -й вершины входит хотя бы одна дуга из указанных в начале абзаца оценок вероятности $p_{i_1 i_2}^{M'_n}$, имеем

$$1 - p_{ii}^{M'_n} \geq c_3 \left[\max_l (1 - W^{(l)}(\mathbf{y}_i)) \right]^n.$$

Подставляя в (8) эту оценку, а также (7) и (9), и пользуясь теоремой 1, связывающей предельные вероятности партий $\pi_n(\mathbf{y})$ с вероятностями $\pi^{M''_n}(s)$, получаем утверждение леммы.

Множество максимальных партий обозначим Y_M , а множество соответствующих вершин графа игры через U_M .

Теорема 3. *Если множество U_M достижимо из любой вершины графа игры, то суммарная предельная вероятность множества максимальных партий стремится к 1 с ростом памяти n :*

$$\lim_{n \rightarrow \infty} \sum_{\mathbf{y} \in Y_M} \pi_n(\mathbf{y}) = 1.$$

Доказательство. Пусть \mathbf{y}_i — какая-нибудь не максимальная партия. По условию существует такая максимальная партия \mathbf{y}_j , что j -я вершина достижима из i -й. Имеем $\max_l (1 - W^{(l)}(\mathbf{y}_i)) > \max_l (1 - W^{(l)}(\mathbf{y}_j))$. Из леммы находим $\lim_{n \rightarrow \infty} \pi_n(\mathbf{y}_i) = 0$, т. е. предельные ве-

роятности всех не максиминных партий стремятся к 0 при росте памяти. Отсюда следует утверждение теоремы.

Использованная в доказательстве лемма показывает, что вероятности $\pi_n(\mathbf{y})$ сходятся к нулю для не максиминных партий, а сумма вероятностей максиминных партий к единице с экспоненциальной скоростью. Заметим еще, что отсутствуют суждения о сходимости каждой из вероятностей $\pi_n(\mathbf{y})$ в случае нескольких максиминных партий.

Пример 2. Три автомата $D_{2,n}$ участвуют в игре с платежными функциями

$$W^{(1)}(\mathbf{y}_1) = W^{(2)}(\mathbf{y}_3) = W^{(3)}(\mathbf{y}_5) = 0,7,$$

$$W^{(1)}(\mathbf{y}_2) = W^{(3)}(\mathbf{y}_6) = 0,3, \quad W^{(1)}(\mathbf{y}_3) = W^{(2)}(\mathbf{y}_1) = \\ = W^{(2)}(\mathbf{y}_6) = 0,8,$$

$$W^{(1)}(\mathbf{y}_4) = W^{(2)}(\mathbf{y}_4) = 0,2, \quad W^{(1)}(\mathbf{y}_5) = W^{(2)}(\mathbf{y}_5) = \\ = W^{(2)}(\mathbf{y}_8) = 0,6,$$

$$W^{(1)}(\mathbf{y}_6) = W^{(3)}(\mathbf{y}_7) = 0,75, \quad W^{(1)}(\mathbf{y}_7) = W^{(3)}(\mathbf{y}_1) = \\ = W^{(3)}(\mathbf{y}_8) = 0,5,$$

$$W^{(1)}(\mathbf{y}_8) = W^{(2)}(\mathbf{y}_7) = W^{(3)}(\mathbf{y}_4) = 0,9,$$

$$W^{(2)}(\mathbf{y}_2) = W^{(3)}(\mathbf{y}_2) = W^{(3)}(\mathbf{y}_3) = 0,4.$$

Множество Y_M состоит из единственной максиминной партии \mathbf{y}_5 . Граф игры изображен на рис. 10. Вершина 5 достижима из всех остальных. Значит, предельная вероятность партии (2, 1, 1) стремится к 1 при $n \rightarrow \infty$. При больших n предельные средние выигрыши автоматов близки к 0,6 для первого и второго участков и к 0,7 для третьего.

Наглядный смысл теоремы 3 опишем, пользуясь антропоморфической терминологией. В условиях этой теоремы автоматы стремятся помочь тому участнику игры, который получает меньше всех, они как бы жертвуют своими доходами, чтобы компенсировать проигравшего.

Какова общность полученного результата? Верно ли, что во всех играх все типы автоматов преимущественно разыгрывают максиминные партии? На этот вопрос дается отрицательный ответ. Можно построить игру с двумя

участниками, в которой автоматы $K_{2,n}$ с гистерезисными переходами (см. § 3 гл. III) «предпочитают» не максминную партию.

Обратимся теперь к одному любопытному классу игр.

Игра с общей кассой называется игра, в которой платежные функции всех участников одинаковы, т. е. при любом y

$$W^{(l)}(y) = W(y).$$

В таких играх все участники в результате каждой партии получают одинаковый средний доход и поэтому их «интересы» совпадают. Они должны стремиться разыгрывать партию-решение, которая доставляет максимум платежной функции. Спрашивается, каковы свойства автоматов в таких играх?

Рассмотрим игру с общей кассой, и пусть в ней участвуют v одинаковых автоматов $D_{k,n}$. Их предельные средние выигрыши одинаковы и обозначаются W_n .

Теорема 4. В игре с общей кассой $\lim_{n \rightarrow \infty} W_n = \max W(y)$.

Доказательство. Для каждой партии игры все автоматы критические. Поэтому на графе игры из любой вершины i исходят дуги во все вершины, которые отвечают партиям, отличающимся от y_i , действием одного автомата. Такие графы сильно связаны, и по теореме 3 вероятность множества Y_M максминных партий с ростом памяти n стремится к 1. На этом множестве функция $W(y)$ достигает максимума.

Примером игры с общей кассой служит игра Гура, определяемая следующим образом. Участники игры могут совершать лишь два действия y_1 и y_2 . Обозначим через m количество участников, избранных первый ход. Платежная функция, одинаковая для всех участников игры, имеет вид $W(m/v)$, т. е. зависит от доли $\theta = m/v$. Эта игра имеет максминные партии-решения, которыми

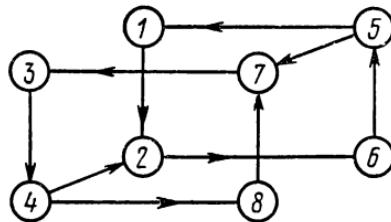


Рис. 10.

являются все партии, в которых действие y_1 совершают $m_0 = \theta_0 v$ участников, где θ_0 определено равенствами $W(\theta_0) = \max_{\theta} W(\theta)$.

По теореме 4 «коллектив» из v автоматов $D_{2,n}$ при достаточно больших n получает в среднем доход, близкий к максимальному.

В заключение этого параграфа рассмотрим свойства автоматов $D_{k,n}$ в играх в размещения. Напомним, что в таких играх партии-ситуации равновесия максминные, но обратное неверно: могут быть максминные партии, не являющиеся ситуациями равновесия.

Теорема 5. *Если в игре в размещения участвуют одинаковые автоматы $D_{k,n}$, то суммарная предельная вероятность множества максминных партий стремится при $n \rightarrow \infty$ к единице.*

Доказательство. Покажем, что всякая игра в размещения удовлетворяет условиям теоремы 3. Для этого достаточно показать достижимость множества U_m из любой вершины, не лежащей в U_m . Пусть $y_i \notin U_m$ и y_m — произвольная максминная партия. Предположим, что для y_i критическими являются автоматы, совершающие действие y_g . Согласно определению максминных партий, в партии y_m либо действие y_g не используется, либо выигрыш автоматов, использующих его, больше, чем у автоматов, выбирающих это действие в партии y_i :

$$\frac{a_g}{v_g(y_m)} \geq \min_j W^{(j)}(y_m) > \min_j W^{(j)}(y_i) = \frac{a_g}{v_g(y_i)}.$$

В обоих случаях $v_g(y_m) < v_g(y_i)$, т. е. действие y_g в партии y_m избирают меньшее число автоматов, чем в партии y_i . Тогда существует h такое, что $v_h(y_m) > v_h(y_i)$, ибо $\sum_{j=1}^k v_j(y_j) = \sum_{j=1}^k v_j(y_m)$. Пусть A_i — один из автоматов, выбирающих в партии y_i действие y_g . Обозначим через y_j партию, полученную из y_i заменой действия автомата A_i с y_g на y_h . По определению графа игры из вершины i существует дуга в вершину j .

Если $y_j \in Y_m$, то достижимость U_m из вершины i доказана, в противном случае рассуждение, проведенное для партии y_i , повторяем для y_j и т. д. В результате получим последовательность вершин i, j, \dots , связанных дугами. Заметим, что

$$\sum_{q=1}^k |\nu_q(y_j) - \nu_q(y_m)| < \sum_{q=1}^k |\nu_q(y_i) - \nu_q(y_m)|$$

и на всех последующих шагах эта сумма также будет уменьшаться, поэтому каждая вершина может встретиться в последовательности не более одного раза, следовательно, через конечное число шагов попадем в множество U_m . Теорема доказана.

§ 3. Игры оценивающих автоматов

Здесь рассматриваются игры, в которых от одного из участников (для определенности будем его считать первым) требуется получать выигрыш, не меньший гарантированного, т. е. не меньше величины

$$v_G = \max_{y'} \min_{y^1, y^2, \dots, y^v} W^{(1)}(y^1, y^2, \dots, y^v).$$

Допустим, что платежная функция $W^{(1)}(y)$ означает не математическое ожидание первого участника, а детерминированное значение его дохода в партии y . Примем в качестве первого участника оценивающий автомат $O_{k, \alpha}$.

Дана игра ν лиц Γ , в которой первым участником является автомат O_{k, α_n} из последовательности автоматов, отвечающей стремящейся к нулю последовательности $\alpha_n \in \mathbb{E}(0, 1)$, а остальными — произвольные конечные вероятностные автоматы $A_n^{(2)}, \dots, A_n^{(v)}$, быть может, принадлежащие ϵ -оптимальным семействам. Предположим, что ассоциированная марковская цепь $M_n = (S^{(n)}, \mathcal{P}^{(n)})$, где $S^{(n)} = S_n^{(1)} \times \dots \times S_n^{(v)}$ — произведение множеств состояний автоматов, а $\mathcal{P}^{(n)}$ — соответствующая матрица вероятностей переходов, является эргодической и поэтому существуют предельные вероятности состояний $\pi^{M_n}(s)$, $s \in S^{(n)}$, не зависящие от начального состояния цепи.

Из высказанных предположений вытекает существование предельного среднего выигрыша W_n автомата $O_{k,n}$. Немаловажно обратить внимание на то, что различным n соответствуют не только разные автоматы $O_{k,n}$, но и, возможно, другие участники игры. Отсюда следует, что множества состояний $S^{(n)}$ марковских цепей M_n с ростом n могут быть различными. При участии в игре бинарных автоматов $A_{k,n}$ из ϵ -оптимальных семейств число состояний в $S^{(n)}$ неограниченно увеличивается. К определенной здесь ситуации относится

Теорема 1. Для предельных средних выигрышей оценивающих автоматов $O_{k,n}$ справедливо неравенство

$$\lim_{n \rightarrow \infty} W_n \geq v_\Gamma.$$

Доказательство основано на исследовании ассоциированной марковской цепи $M_n = (S^{(n)}, \mathcal{P}^{(n)})$. Состояния ее по-прежнему обозначаем через $s = (s', \dots, s^n)$.

Обозначим выигрыш, получаемый оценивающим автоматом, когда система находится в состоянии s , через $W(s)$. Утверждение теоремы вытекает из следующего: суммарная предельная вероятность состояний, для которых $W(s) < v_\Gamma$, стремится к нулю при $n \rightarrow \infty$.

Разобьем множество $S^{(n)}$ на k подмножеств $S_1^{(n)}, \dots, S_k^{(n)}$ следующим образом: $s \in S_i^{(n)}$, если первая компонента вектора s (т. е. состояние оценивающего автомата) есть s_i . Пусть при действии y , минимальный выигрыш первого автомата равен гарантированному v_Γ , т. е. $\min_{s \in S_r^{(n)}} W(s) = v_\Gamma$, а при действии y_q он меньше чем v_Γ . Обозначим через $F_q^{(n)}$ подмножество тех состояний $S_q^{(n)}$, для которых $W(s) < v_\Gamma$, и через

$$\bar{W} = \max_{s \in F_q^{(n)}} W(s) < v_\Gamma.$$

Зафиксируем произвольное $\epsilon > 0$ и покажем, что при всех n , превосходящих некоторое n_0 , будет $\sum_{s \in F_q^{(n)}} \pi^{M_n}(s) < \epsilon$.

Рассмотрим цепи M'_n , которые имеют k состояний и переходные вероятности которых вычисляются по

формуле

$$p_{ij}^{M'_n} = \sum_{\bar{s} \in S_i^{(n)}} \frac{\pi^{M_n}(\bar{s})}{\sum_{\bar{s}' \in S_i^{(n)}} \pi^{M_n}(\bar{s}')} p_{\bar{s}S_j^{(n)}}, \quad (1)$$

где $p_{\bar{s}S_j^{(n)}}^{M_n} = \sum_{\bar{s}' \in S_j^{(n)}} p_{\bar{s}\bar{s}'}$. Непосредственно проверяется, что предельная вероятность i -го состояния цепи M'_n равна суммарной предельной вероятности множества состояний $S_i^{(n)}$ в цепи M_n . Легко убедиться также в справедливости соотношений

$$\frac{\pi^{M'_n}(s_i)}{\pi^{M'_n}(s_j)} = \frac{p_{ji}^{M'_n}}{p_{ij}^{M'_n}} \quad (2)$$

и

$$p_{ij}^{M'_n} \leq \max_{\bar{s} \in S_i^{(n)}} p_{\bar{s}S_j^{(n)}}^{M_n}. \quad (3)$$

Оценим переходные вероятности $p_{rq}^{M'_n}$, $p_{qr}^{M'_n}$. Нетрудно видеть, что $p_{\bar{s}S_l^{(n)}}^{M_n} = \frac{g_{\alpha_n}(W(\bar{s}))}{k-1}$, поэтому в силу (3) имеем

$$p_{rq}^{M'_n} \leq \frac{g_{\alpha_n}(v_\Gamma)}{k-1}.$$

Объединим в множество I_1 те значения индекса n , при которых

$$\sum_{\bar{s} \in F_q^{(n)}} \pi^{M_n}(\bar{s}) \geq \epsilon \sum_{\bar{s}' \in S_q^{(n)}} \pi^{M_n}(\bar{s}'),$$

и в множество I_2 — остальные значения n . Для $n \in I_2$ сразу имеем $\sum_{\bar{s} \in F_q^{(n)}} \pi^{M_n}(\bar{s}) < \epsilon$, а для $n \in I_1$ из (1) получаем оценку

$$p_{qr}^{M'_n} \geq \frac{\epsilon}{k-1} g_{\alpha_n}(\bar{W}).$$

Подставляя в (2), находим

$$\frac{\pi^{M'_n}(s_q)}{\pi^{M'_n}(s_r)} \leq \frac{g_{\alpha_n}(v_\Gamma)}{\epsilon g_{\alpha_n}(\bar{W})}.$$

При всех n , превосходящих некоторое n_0 , $g_{\alpha_n}(v_\Gamma) < \epsilon^2 g_{\alpha_n}(\bar{W})$.

Таким образом, для $n \in I_1$, $n > n_0$

$$\sum_{s \in S_q^{(n)}} \pi^{M_n}(s) = \pi^{M'_n}(s_q) < \varepsilon.$$

Теорема доказана.

Перейдем к рассмотрению игр оценивающих автоматов между собой. Пусть задана игра Γ и ν последовательностей оценивающих автоматов. Дополнительно будем предполагать, что эти автоматы «не сильно отличаются» между собой, а именно, что для любой пары i, j и любого n выполняется неравенство

$$\frac{g_{\alpha_n}^i(a_1)}{g_{\alpha_n}^j(a_2)} < \alpha_n, \quad \text{если } a_1 \in \Delta_i, a_2 \in \Delta_j, a_1 > a_2, \quad (4)$$

и

$$\frac{g_{\alpha_n}^i(a)}{g_{\alpha_n}^j(a)} > c, \quad \text{если } a \in \Delta_i \cap \Delta_j, \quad (5)$$

где $c > 0$ — некоторая константа.

В цепях M_n , описывающих игру автоматов, всякой партии соответствует одно состояние и вероятность партии определяется как предельная вероятность этого состояния.

Оказывается, что на игры оценивающих автоматов можно перенести большую часть результатов, полученных для игр бинарных автоматов из § 2.

Сопоставим игре Γ ориентированный граф V_Γ , построенный по платежным функциям точно таким же образом, как в § 2. Как и там, Y_M означает множество максиминных партий и U_M — соответствующее множество вершин графа V_Γ .

Теорема 2. *Если множество U_M достижимо из любой вершины графа игры, то суммарная предельная вероятность множества максиминных партий стремится к единице при $n \rightarrow \infty$.*

Доказательство. Введем обозначение: $W(y) = \min_{1 \leq i \leq \nu} W^{(i)}(y)$ для минимального выигрыша в партии y . Достаточно показать, что если вершина l дости-

жима из вершин k и $W(\mathbf{y}_i) > W(\mathbf{y}_j)$, то предельная вероятность партии \mathbf{y}_k стремится к нулю.

Рассмотрим последовательность цепей M'_n такую, что переходные вероятности M'_n и M_n связаны таким же образом, как переходные вероятности M''_n и M'_n в § 2 (соотношение (7')). Предельные вероятности состояний в цепях M_n и M'_n связаны соотношением

$$\frac{\pi^{M_n}(s_j)}{\pi^{M_n}(s_i)} = \frac{\pi^{M'_n}(s_j)(1 - p_{jj}^{M_n})^{-1}}{\pi^{M'_n}(s_i)(1 - p_{ii}^{M_n})^{-1}}. \quad (6)$$

Оценим сначала величину $1 - p_{jj}^{M_n}$. Пусть μ_k обозначает номер автомата, получающего в партии \mathbf{y}_j минимальный выигрыш $W(\mathbf{y}_j)$ (если таких автоматов несколько, берем в качестве μ_j номер любого из них). Тогда

$$1 - p_{jj}^{M_n} = \sum_{l \neq j} p_{jl}^{M_n} > g_{\alpha_n}^{\mu_j}[W(\mathbf{y}_j)],$$

$$1 - p_{jj}^{M_n} < g_{\alpha_n}^1[W^{(1)}(\mathbf{y}_j)] + \dots + g_{\alpha_n}^v[W^{(v)}(\mathbf{y}_j)] \leqslant$$

$$\leqslant \frac{v}{c} g_{\alpha_n}^{\mu_j}[W(\mathbf{y}_j)], \quad (7)$$

где c взято из соотношения (5).

Если из вершины j_1 существует дуга в вершину j_2 , то партии \mathbf{y}_{j_1} и \mathbf{y}_{j_2} по определению отличаются действием одного (m -го) автомата и $W^{(m)}(\mathbf{y}_{j_1}) = W(\mathbf{y}_{j_1})$, следовательно,

$$p_{j_1 j_2}^{M_n^2} = \frac{g_{\alpha_n}^m[W(\mathbf{y}_{j_1})]}{(k-1)(1 - p_{j_1 j_2}^{M_n})} \geqslant \frac{cg_{\alpha_n}^m[W(\mathbf{y}_{j_1})]}{(k-1)v g_{\alpha_n}^{\mu_j}[W(\mathbf{y}_{j_1})]} > \frac{c^2}{(k-1)v}.$$

Отсюда так же, как в доказательстве теоремы 3 из § 2, показывается, что

$$\pi^{M'_n}(s_i) > \left[\frac{c^2}{(k-1)v} \right]^R \pi^{M'_n}(s_j).$$

Подставляя найденные оценки в (6), получаем

$$\frac{\pi^{M_n}(s_j)}{\pi^{M_n}(s_l)} < \left[\frac{(k-1)v}{c^2} \right]^R \frac{v g_{\alpha_n}^{u_l}[W(y_l)]}{c g_{\alpha_n}^{u_j}[W(y_j)]}.$$

По предположению $W(y_l) > W(y_j)$, следовательно, из (4) вытекает $\pi^{M_n}(s_j) \rightarrow 0$ при $n \rightarrow \infty$. Теорема доказана.

Доказательство теорем 4 и 5, которые в § 2 описывали поведение бинарных автоматов в играх с общей кассой и в играх в размещения, основывалось лишь на теореме 3, поэтому доказанная только что теорема позволяет сразу перенести эти результаты на игры оценивающих автоматов с общей кассой и в размещения. Мы не будем останавливаться на этих вопросах.

§ 4. Игры автоматов. Моделирование

Результаты § 2 об играх бинарных автоматов (точнее говоря, автоматов $D_{k,n}$) относились к асимптотическому поведению при неограниченном росте памяти n . Числовые характеристики конкретных игр там получить не удалось. Поэтому остаются неясными даже простейшие свойства автоматов. Например, при фиксированном n нас интересует: во-первых, каковы предельные вероятности партий в данной игре; во-вторых, чему равны средние выигрыши участников игры и т. д. По-видимому, аналитические методы не могут пока дать ответ и единственным эффективным средством вычисления этих характеристик остается моделирование игры на ЦВМ. Для этого сначала задаются платежные функции и конструкции автоматов-участников. Обычные процедуры метода Монте-Карло дают исходы партий, по которым подсчитываются текущие величины (на момент t) эмпирических средних выигрышей автоматов

$$V_t^{(j)} = \frac{1}{t} \sum_{l=1}^t x_l^{(j)}, \text{ где } x_l^{(j)} \text{ — последовательность значений}$$

j -й компоненты управляемого ОПНЗ, поступающих на вход j -го автомата. Кроме того, подсчитываются частоты наступления интересующих партий. В случае эргодических игр автоматов (им отвечает эргодическая ассоции-

рованная марковская цепь) величины $V_i^{(j)}$ и частоты партий оказываются близкими соответственно к $W_n^{(j)}$ и $\pi_n(y)$ при данной величине памяти n .

В этом параграфе изложены некоторые численные результаты иллюстративного характера о поведении бинарных автоматов и δ -автоматов в играх.

Начнем с бинарных автоматов.

В качестве участников игр изберем автоматы с линейной тактикой $L_{k,n}$. Их достоинством при моделировании является небольшая, по сравнению с автоматами $D_{k,n}$, «инерционность», т. е. умеренная продолжительность совершения действий. Это обстоятельство сокращает время моделирования, однако даже в этих условиях количество партий, за которое оценивают предельные вероятности, достигает величины 10^5 . Спрашивается, каковы связи теорем § 2 об автоматах $D_{k,n}$ с численными данными об автоматах $L_{k,n}$? Есть ли смысл изучать свойства $L_{k,n}$ в конкретных играх, если общие факты доказаны для других автоматов? Ответ на этот вопрос утвердительный, ибо можно показать, что в предположении $\max W^{(j)}(y) > 1/2$ результаты § 2 сохраняют силу.

Мы снова сопоставим наказанию автомата число -1 , а не 0 , как в § 2. По-прежнему поощрению отвечает число 1 . Вероятности поощрения q и наказания p выражаются через платежную функцию W уже известными нам формулами

$$q = \frac{1 + W}{2}, \quad p = \frac{1 - W}{2}.$$

Рассмотрим игру с двумя участниками, имеющими по два действия y_1 и y_2 , определяемую платежной функцией $W(y', y'')$. Она означает средний выигрыш того участника, который избрал y' , в то время как соперник совершил действие y'' . Значения W примем такими:

$$W(y_1, y_1) = W(y_2, y_2) = 0,25$$

$$W(y_1, y_2) = W(y_2, y_1) = 0,1.$$

Партия $y=(y_1, y_2)$ является и ситуацией равновесия, и максиминной партией. При смене каждым игроком действия в этой партии он терпит большой убыток. Частоты пар-

тий приведены в табл. 6.1. Из нее мы видим, что с увеличением памяти автоматов n частота партии $y = (y_1, y_1)$ приближается к 1 и уже при $n=6$ близка к 1.

Таблица 6.1

y	n		
	4	6	8
(y_1, y_1)	0,72	0,86	0,94
$\begin{cases} (y_1, y_2) \\ (y_2, y_1) \end{cases}$	0,26	0,14	0,06
(y_2, y_2)	0,01	0	0

Участники игры на окружности имеют по одному и тому же числу действий k . Платежная функция j -го участника зависит от его действия и, в простейшем случае, от действий его «соседей» слева и справа, т. е. участников с номерами $j-1$ и $j+1$ (подразумевается, что $(v+1)$ -й игрок отождествлен с 1-м, а 0-й с v -м). Это означает, что в партии $y = (y_{l_1}, \dots, y_{l_v})$ средний выигрыш j -го участника равен $W^{(j)}(y_{l_{j-1}}, y_{l_j}, y_{l_{j+1}})$. Пусть функция $W^{(j)}$ не зависит от j . Рассмотрим пример игры на окружности, заданной такими значениями платежной функции (при $k=2$):

$$W(y_1, y_1, y_1) = 0,43, \quad W(y_2, y_2, y_2) = 0,6,$$

а для всех остальных троек аргументов — нули. Партия $y' = (y_1, \dots, y_1)$ есть ситуация равновесия, а $y'' = (y_2, \dots, y_2)$ является решением и максминной. В табл. 6.2 указаны значения доли партии y'' после 10^5 партий, разыгранных автоматами $L_{2,n}$.

Мы видим, что автоматы $L_{2,n}$ уже при $n=7$ подавляющее время реализуют партию y'' . Увеличение количества участников игры v уменьшает (видимо, степенным образом) при фиксированном n эту долю.

Видоизменим платежную функцию. Теперь пусть $W(y_1, y_1, y_2) = W(y_2, y_1, y_1) = 0,2, \quad W(y_1, y_2, y_1) = 0,$

$W(y_1, y_2, y_2) = W(y_2, y_2, y_1) = -0,2$, $W(y_2, y_1, y_2) = 0$
и по-прежнему $W(y_1, y_1, y_1) = 0,43$, $W(y_2, y_2, y_2) = 0,6$.

Таблица 6.2

n	v			
	3	9	18	32
3	0,54	0,22	0,03	0,00
5	0,91	0,78	0,60	0,45
7	0,99	0,98	0,91	0,81

В этих условиях максиминная партия-решение y'' оказывается «неустойчивой»: стоит одному из партнеров заменить действие y_2 на y_1 , как его соседи начинают проигрываясь (наказываться) с большой вероятностью, и поэтому автоматы $L_{2,n}$ сменяют действие. Это соображение подтверждается результатами моделирования (табл. 6.3), в которой даны частоты партии y' . Таким образом, максиминная партия y'' не является предельной в этой игре автоматов $L_{2,n}$. Нетрудно проверить, что на графе игры эта партия недостижимая, и воспользоваться результатами теоремы 3 из § 2 нельзя.

Обратимся к свойствам автоматов $L_{k,n}$ в играх в размещения, которые определены в конце § 1. В двух рассматриваемых примерах полагаем, что в игре участвуют по $v=5$ автоматов $L_{7,n}$.

Сначала игру зададим числами $a_1=0,9 \cdot a_2=\dots=a_7=0,33$. Ситуация равновесия и максиминные партии отвечают такому размещению автоматов по действиям: $(2, 1, 1, 1, 0, 0, 0)$. Моделирование для глубины памяти $n=5$ привело к следующему результату: действие y_1 «занято» одним автоматом в 35% партий, двумя автоматами — в 55% и тремя — в 10%. Остальные действия в 4% партий совершались одновременно двумя автоматами. Рост памяти автоматов улучшает их «поведение» в игре, а именно при $n=10$ действие y_1 один автомат избирает в 9% партий, два автомата — в 79% и три — в 12%. Остальные действия всегда совершаются одним автоматом.

Снова мы видим, что автоматы $L_{2,n}$ при умеренной глубине памяти демонстрируют поведение, которое предписывает теория.

Таблица 6.3

n	ν		
	6	18	32
5	0,79	0,42	—
6	0,88	0,69	—
7	0,92	0,82	0,69

Таблица 6.4

n	5	10	20
	0,25	0,23	0,21

Пусть все числа a_i одинаковые: $a_1 = \dots = a_7 = 0,6$. Оказалось, что при $n=6$ в 97% партий каждое действие совершает один автомат.

Наконец, примем $a_1=0,9$, $a_2=\dots=a_7=0,15$. Здесь максминной партией (и ситуацией равновесия) служит выбор всеми автоматами первого действия, приводящий к среднему выигрышу 0,18. Однако автоматы добиваются большего выигрыша, совершая u_1 вчетвером, иногда втроем и даже вдвоем. Состав этой группы меняется, и в результате получаемый ими средний выигрыш превышает теоретическое значение 0,18. В табл. 6.4 приведены величины средних выигрышей автоматов $L_{7,n}$ при различных n . Из этих данных видно, что средние выигрыши автоматов с ростом n приближаются к величине 0,18.

Перейдем к числовым характеристикам поведения автоматов $L_{2,n}$ в игре Гура (см. § 2). Эти характеристики зависят не только от глубины памяти n , но и от числа ν принимающих участие в игре автоматов. В экспериментах на ЦВМ была принята платежная функция (с аргументом θ , означающим долю автоматов, которые совершили первое действие) такого вида: на отрезке $[0, 1; 0, 2]$ линейно растет от 0,2 до 0,8, на отрезке $[0, 2; 0, 3]$ линейно убывает от 0,8 до 0,2 и на остальной части отрезка $[0, 1]$ постоянна и равна 0,2. Моделирование подтвердило, что действительно реализуются заранее предсказуемые возможности: при фиксированном n (принимаемом при

имитации небольшим) и растущем числе автоматов v «коллектив» автоматов разбивается на две равные по численности группы, одна из которых совершают действие y_1 , а другая — y_2 ; при фиксированном числе участников увеличение глубины памяти вызывает увеличение частоты партий-решений, соответствующих максимуму платежной функции (равному 0,8 в точке $\theta=0,2$). Среди числовых результатов интересно отметить такой: при $v=10$ и $n=5$ частота партии-решения оказывается весьма близкой к единице (после 10^5 партий игры). Уменьшение глубины памяти до 4 и до 3 снижает эту частоту до 0,7 и 0,2 соответственно. При этом возрастает доля партий, в которых $m/v \approx 0,5$. Заметим, что при увеличении числа участников игры для сохранения близких к единице частот партий-решения необходимо увеличивать глубину памяти автомата. Это обстоятельство было уже замечено в играх на окружности.

Приведенные выше численные результаты моделирования игр бинарных автоматов являются типическими. В многочисленных сериях экспериментов получены сходные факты. Резюмируются они следующим образом:

1) Средние выигрыши автоматов и частоты партий делаются стабильными при числе сыгранных партий порядка 10^5 .

2) Значения средних выигрышей автоматов и частот максимальных партий оказываются близкими к предельным теоретическим при умеренной (даже небольшой) глубине памяти, равной 6—10 (кроме игры в размещении автоматов $L_{k,n}$, где n должно быть большим).

Сделаем теперь краткий обзор итогов имитации игр δ -автоматов, снова ограничиваясь малым числом характерных примеров.

Осмысливание результатов моделирования игр δ -автоматов затруднено тем обстоятельством, что неизвестны цели, достижаемые «коллективом» таких автоматов при управлении векторными ОПНЗ. Поэтому характерными партиями будем считать ситуации равновесия и сопоставлять эмпирические доходы автоматов с доходами в этих партиях.

Во всех играх рассматривались автоматы типа G , точнее, автоматы «с забыванием» G_k . Их параметры, общие

для всех рассмотренных игр, таковы: количества действий k от 2 до 7, $\theta=0,5$, $\hat{p}=0,8$, вектор вероятностей действий p преобразуется на каждом такте, начиная с k -го, после начала функционирования автомата. Значения δ укажем для каждой игры отдельно. В качестве способа «забывания» изберем $\Phi_{2,b}$ (см. § 4 гл. IV).

Компоненты управляемого процесса (v -мерного ОПНЗ) имеют вид

$$\xi^{(j)}(y) = \zeta^{(j)}(y) + W^{(j)}(y),$$

где $\zeta^{(j)}(y)$ — случайная величина с нулевым математическим ожиданием и вероятностями значений $P_j(x|y)$. В ходе имитации все $\xi^{(j)}$ одинаково распределены — равномерно на множестве двоично-рациональных чисел интервала $(-1, 1)$, представимых в разрядной сетке ЦВМ.

Критерием эффективности поведения автоматов в играх служат величины эмпирических средних выигрышей

$$V_T^{(j)} = \frac{1}{T} \sum_{i=1}^T x_i^{(j)},$$

где T должно быть достаточно большим, чтобы парировать случайные отклонения. Мы не знаем, существуют ли пределы $V^{(j)} = \lim_{T \rightarrow \infty} V_T^{(j)}$ для игр автоматов типа G . Эксперименты всегда свидетельствуют в пользу положительного ответа: начиная с $T=1,5 \cdot 10^3 \div 2,5 \cdot 10^3$ средние выигрыши оставались практически неизменными (всего игры продолжались $10^4 \div 10^5$ партий).

Ниже приводятся результаты исследования таких игр, в которых заранее естественно ожидать близость всех $V_T^{(j)}$. Это предположение подтвердилось, и поэтому за количественную характеристику поведения автоматов в игре удобно принять «среднюю» цену

$$V_T = \frac{1}{v} \sum_{j=1}^v V_T^{(j)}.$$

Рассмотрим игру в размещения, задаваемую числами

$$a_1 = 5, a_2 = \dots = a_7 = 2.$$

Участники игры — 5 автоматов G_n — имеют по 7 действий и значение $\delta=0,1$. На рис. 11 приведена зависимость V_t от параметра памяти h , которая носит монотонный характер. Это факт легко понять: действия «партнеров» по игре делают управляемый процесс для каждого автомата неоднородным, и отыскание лучших действий в данный момент требует забывания автоматом откликов, полученных за давно совершенные действия. В нашем примере цена партии-ситуации равновесия (т. е. $W(y_R)=$

$$=\frac{1}{v} \sum_1^v W^{(j)}(y_R) \text{ равна}$$

2,20. Моделирование с $h=10$ привело после $1,8 \cdot 10^3$ партий к средней цене $V=2,15$, т. е. «потери» автоматов составили всего менее 2,5%.

Уменьшим величину δ . Будем ее полагать (всюду до конца этого параграфа) равной $\delta=0,03$. Введем, кроме предыдущего варианта, обозначаемого Γ_1 , еще два варианта игры в размещения:

$$\Gamma_2: \quad a_1 = 9, a_2 = \dots = a_7 = 0,33,$$

$$\Gamma_3: \quad a_1 = 9, a_2 = \dots = a_7 = 1,5.$$

Результаты моделирования этих игр сведены в табл. 6.5. В ее первой строке указаны цены ситуаций равновесия, а во второй — эмпирические цены этих партий (после 10^4 сыгранных партий). Из данных мы замечаем, что в играх Γ_1 и Γ_2 автоматы G_h достигают цен партий, превосходящих цену ситуации равновесия, а в игре Γ_3 доходы автомата практически совпадают с доходами в партии-ситуации равновесия.

Обратимся в заключение к поведению автоматов G_h в игре Гура. Теперь автоматы в количестве 10 имеют по 2 действия.

Эксперименты вновь показали, что средние выигрыши V_t монотонно убывают с ростом параметра h . Для пла-

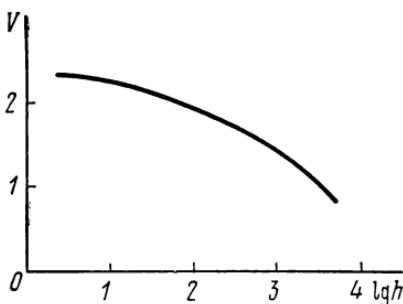


Рис. 11.

тежной функции $W(\theta)$, аналогичной по виду той, которая использовалась в играх бинарных автоматов, но со значением максимума в точке $\theta=0,3$, равным 5, результаты

Таблица 6.5

g_1	g_2	g_3
2,20	3,76	3,00
2,42	3,92	2,98

следующие. В 92% партий, реализуется партия-решение, т. е. 3 автомата совершают первое действие. Средняя цена составляет 4,92, т. е. «потери» автоматов оказались почти 1,5%. В случае нескольких равных максимумов платежной функции, частоты всех партий-решений одинаковы и их сумма близка к единице.

Отметим, что «коллектив» автоматов G_h выходит на партию-решение в среднем за $1,8 \cdot 10^3$ партий. После этого средние выигрыши автоматов остаются практически неизменными.

§ 5. Адаптивные иерархические системы

В предыдущих параграфах рассматривалась простейшая схема управления — децентрализованная — в виде прямого произведения элементарных систем, конечных автоматов. Возможности систем такого типа ограничены. Поэтому часто оказывается необходимым использовать «сложные» системы. Сложность управляющей системы заключается в том, что она состоит из многих взаимосвязанных подсистем. Подчеркивая важность взаимозависимости подсистем, будем говорить об иерархических системах. Определим это понятие, не ограничивая объекты управления лишь управляемыми случайными процессами типа ОПНЗ.

Задан класс K управляемых случайных процессов ξ_t , с фазовым пространством X и пространством управлений Y . Цель управления относится к системе функционалов $(\varphi_t^{(1)}, \dots, \varphi_t^{(m)})$, определенных на ξ_t . Сами функционалы выбраны из класса Φ допустимых функционалов.

Введем основную управляющую систему U_o , которая представляет собой объект вида

$$U_o = [Z, D, T, Y],$$

где Z — множество входных сигналов (это либо значения процесса ξ_t , — точки фазового пространства X , либо значения набора функционалов $\varphi_t^{(1)}, \dots, \varphi_t^{(m)}$), Y — множество выходных сигналов, D — множество правил выбора действий (по запоминаемому прошлому) и T — семейство операторов на D , определяющих используемое правило из множества D . Пара (D, T) представляет собой стратегию для достижения цели, относящейся к процессу из класса K и набору заданных на нем функционалов.

Пусть задано пространство дополнительных управлений $\tilde{Y} = \tilde{Y}_1 \times \dots \times \tilde{Y}_L$, от компонент которого управляемый процесс не зависит.

Введем дополнительные управляющие системы

$$U_\lambda = [Z_\lambda, D_\lambda, T_\lambda, \tilde{Y}_\lambda], \quad \lambda = 1, \dots, L,$$

у которых символы в квадратных скобках имеют смысл, аналогичный тем, которые фигурируют в обозначении U_o . Отличие U_λ от U_o состоит в множествах входных сигналов. Множествами Z_λ служат наборы значений функционалов в количестве m_λ из семейства $\{\varphi_t^{(1)}, \dots, \varphi_t^{(N)}\}$, $m_\lambda \leq N$, каждый из которых зависит от управляемого процесса ξ_t и некоторых компонент вектора дополнительных управлений $y = (\tilde{y}^{(1)}, \dots, \tilde{y}^{(L)}) \in \tilde{Y}$. По отношению к этим m_λ функционалам система U_λ является стратегией.

Скажем, что управляющая система U_λ , подчинена управляющей системе $U_{\lambda''}$, если реализуемая системой U_λ стратегия явным образом зависит от дополнительных управлений $\tilde{y}^{(\lambda'')} \in \tilde{Y}_{\lambda''}$. Система U_λ , называется *подчиненной* системой, а $U_{\lambda''}$ *подчиняющей*. Строго говоря, выходные сигналы подчиняющей системы должны считаться входными сигналами подчиненной системы, т. е. их следовало бы явно включить в множество Z_λ . Мы этого не делаем.

Введенное бинарное отношение делает множество дополнительных управляющих систем $\{U_\lambda\}$ частично упорядоченным. Обозначим отношение подчинения через Π . Допустим еще, что система U_o подчинена некоторым (в частности, всем) дополнительным системам U_λ .

Иерархической управляющей системой называется объект

$$J = [U_o; U_\lambda, \lambda = \overline{1, L}; \Pi].$$

Иерархическую управляющую систему можно изобразить ориентированным графом, вершинами которого служат системы U_o , U_λ , а дуги выходят из вершин, соответствующих подчиняющим системам, в вершины, отвечающие подчиненным системам. Особую роль играют иерархические системы с графиками в виде деревьев.

В каждый момент функционирования иерархической системы система U_o и любая из U_λ действует с теми элементами множеств Z , D , T , которые определяются поступившими на предыдущем такте (или ранее) значениями дополнительных управлений. Мы не предполагаем, что все подсистемы воспринимают входные сигналы и совершают действия одновременно. Подчиняющие системы могут действовать более редко, чем подчиненные.

Иерархические управляющие системы являются частным случаем обучаемых систем и реализуют некоторую стратегию. Свойство адаптивности иерархических систем заключается в обеспечении достижения цели сразу на классе объектов.

Мы здесь рассмотрим лишь один пример адаптивной иерархической системы.

Пусть управляемый случайный процесс ξ_t — склярный ($X = [a, b]$), имеет конечное множество управлений $Y = \{y_1, \dots, y_k\}$ и характеризуется d распределениями $\mu_j(x|y)$, $j = 1, \dots, d$, которые периодически сменяют друг друга, т. е. в моменты времени $t = ld + i$, $l = 0, 1, 2, \dots$, процесс имеет распределение μ_i , $i = 1, \dots, d$. Итак, ξ_t представляет собой неоднородный процесс с независимыми значениями, который естественно назвать *периодическим*. Обозначим через $W_j(y)$ средний выигрыш, отвечающий j -му распределению. Поставим целью управления максимизацию функции

$$W(y^{(1)}, \dots, y^{(d)}) = \frac{1}{d} \sum_{j=1}^d W_j(y^{(j)}). \quad (1)$$

Если все распределения μ_j (либо выигрыши W_j) известны,

решение задачи очевидно: нужно для каждого j найти «оптимальные действия» $y_o^{(j)}$ и их затем циклически применять. Мы предположим, что распределения μ_1, \dots, μ_d неизвестны. Более того, допустим, что неизвестен период d , но указана его верхняя оценка \bar{d} ($d \leq \bar{d}$).

Получим решение задачи адаптивного управления двухуровневой иерархической системы. Ее первым уровнем, основной управляющей системой служат \bar{d} автоматов $A_1, \dots, A_{\bar{d}}$. Ими могут быть либо автоматы из ϵ -оптимального семейства, либо асимптотически оптимальные автоматы. Выбор тех или иных конструкций определяется заданной целью управления — максимизацией функции (1), приближенной (с точностью до ϵ) или же точной. Теперь следует задать правило выделения из \bar{d} имеющихся автоматов (A_i) d автоматов, которые поочередно вступают в управление процессом ξ_t . Для этого введем второй уровень иерархической системы \bar{d} бинарных автоматов $B_1, \dots, B_{\bar{d}}$ с двумя действиями y', y'' каждый, принадлежащих ϵ -оптимальным семействам. Взаимодействие этих автоматов определяется следующей игрой с общей кассой, вариантом игры Гура. Пусть из автоматов $A_1, \dots, A_{\bar{d}}$ ровно l штук ($0 < l \leq \bar{d}$) осуществляли управление процессом ξ_t и получили входные сигналы $x^{(1)}, \dots, x^{(l)}$. Образуем величину

$$P = \frac{1}{(b-a)l} \sum_{j=1}^l (x^{(j)} - a),$$

которую примем за общую для всех автоматов вероятность поощрения $B_1, \dots, B_{\bar{d}}$, причем входные сигналы поступают на них независимым образом и автоматы, сменив состояния, совершают действия. Пусть действие y' избрали автоматы B_{j_1}, \dots, B_{j_l} , тогда на протяжении l тактов процессом ξ_t управляют автоматы A_{j_1}, \dots, A_{j_l} (остальные не взаимодействуют с процессом), «включаемые» последовательно один за другим. Из входных сигналов этих автоматов формируется, как и выше, вероятность поощрения автоматов $\{B_j\}$, и цикл повторяется.

Эта двухуровневая система управления периодическим процессом с независимыми значениями способна обеспе-

чить заданную цель, если совокупность автоматов B_1, \dots, B_d может в указанной разновидности игры Гура устойчиво разделиться на две группы, из которых одна, избирающая действие y' , состоит из кратного d числа автоматов.

Для установления свойств этой управляющей системы примем, что автоматами первого уровня $\{A_j\}$ являются автоматы с линейной тактикой $L_{k,n}$, имеющие по k действий, а автоматами второго уровня $\{B_j\}$ — автоматы с линейной тактикой $L_{2,n}$. Отметим, что у всех автоматов системы глубина памяти n одна и та же. Нас будут интересовать асимптотические свойства системы: способность при большой памяти приближаться к максимуму предельного среднего выигрыша.

Скажем, что система находится в состоянии $\sigma = \sigma(l, i, y)$, где i и y суть l -мерные векторы; если «включены» l автоматов уровня A , номера этих автоматов суть компоненты вектора i , а действия, совершаемые ими, указываются компонентами вектора y .

Через S_0 обозначим множество всех оптимальных состояний, в них число действующих автоматов уровня A равно периоду процесса, а действия этих автоматов наилучшие. Через S_1 обозначим множество остальных (неоптимальных) состояний. В состоянии из S_0 система получает максимально возможный средний выигрыш W_{\max} . Введем еще $W(\sigma)$ — средний выигрыш в состоянии σ . Основной результат относительно рассматриваемой системы сформулирован в следующей теореме ($W_n(t)$ означает средний выигрыш системы в момент времени t).

Теорема. $\lim_{n \rightarrow \infty} \lim_{t \rightarrow \infty} W_n(t) = W_{\max}$.

Доказательство. Полагая $W' = \max_{\sigma \in S_1} W(\sigma)$, $W'' = \min_{\sigma \in S_1} W(\sigma)$, запишем неравенства для среднего выигрыша

$$\frac{1}{t} \mathbf{E} [W_{\max} \tau_t(S_0) + W'' \tau_t(S_1)] \leq W_n(t) \leq \frac{1}{t} \mathbf{E} [W_{\max} \tau_t(S_0) + W' \tau_t(S_1)].$$

Здесь $\tau_t(\cdot)$ — суммарное время, проведенное в множестве состояний (\cdot) на отрезке $[1, t]$. Переходя к пределу сначала по t , а затем по n , и учитывая, что $\tau_t(S_0) + \tau_t(S_1) = t$,

убеждаемся, что для доказательства утверждения теоремы достаточно показать справедливость равенства

$$\lim_{n \rightarrow \infty} \lim_{t \rightarrow \infty} \frac{E\tau_t(S_1)}{E\tau_t(S_0)} = 0. \quad (2)$$

Рассмотрим функционирование системы на отрезке $[1, t]$. Пусть система за данный отрезок времени $\mu_0(t)$ раз побывала в множестве состояний S_0 и $\mu_1(t)$ раз — в множестве состояний S_1 . Ясно, что $|\mu_0(t) - \mu_1(t)| \leq 1$. Пусть также при j -м попадании в множество S_x , $x = 0, 1$, система побывала в состояниях $\sigma_j^1, \dots, \sigma_j^{v_x(j)} \in S_x$. Тогда интересующая нас величина запишется в виде

$$\frac{E\tau_t(S_1)}{E\tau_t(S_0)} = \frac{E \left(\sum_{j=1}^{\mu_1(t)} \sum_{m=1}^{v_1(j)} \tau(\sigma_j^m) \right)}{E \left(\sum_{i=1}^{\mu_0(t)} \sum_{l=1}^{v_0(i)} \tau(\sigma_i^l) \right)}, \quad \sigma_i^l \in S_0, \quad \sigma_j^m \in S_1. \quad (3)$$

Сформулируем теперь вспомогательные предложения.

1. Пусть $\sigma \in S_0$. Тогда

$$E\tau(\sigma) \geq c_1 \left(\frac{q}{1-q} \right)^n = A(n),$$

где $c_1 > 0$ не зависит от n , $q > 1/2$.

2. Пусть $\sigma \in S_1$. Тогда

$$E\tau(\sigma) \leq \left(\frac{q_1}{1-q_1} \right)^n = B(n),$$

где $q_1 < q$.

3. Пусть система находилась в множестве состояний S_1 и за время от момента попадания в S_1 до момента выхода совершила v переходов между состояниями из S_1 . Тогда

$$E\tau \leq c_2 n^\gamma = C(n),$$

где $c_2 > 0$ и $\gamma > 0$ не зависят от n .

С помощью 1—3 из (3) делаем заключение

$$\frac{E\tau_t(S_1)}{E\tau_t(S_0)} \leq \frac{A(n)C(n)}{B(n)} \frac{E\mu_1(t)}{E\mu_0(t)},$$

а отсюда выводим (2).

ГЛАВА VII

УПРАВЛЕНИЕ ОБОБЩЕННЫМИ ПРОЦЕССАМИ С НЕЗАВИСИМЫМИ ЗНАЧЕНИЯМИ

§ 1. Предварительные замечания.

До сих пор мы занимались задачами управления процессами из класса ОПНЗ. Роль таких процессов, характеризуемых вероятностями $\mu(M|y)$, определяется их простотой. Однако теоретические и практические соображения делают необходимым исследование более широких классов процессов. В этой главе рассматриваются классы, которые непосредственно обобщают ОПНЗ.

Управляемый случайный процесс ξ_t с фазовым пространством (X, \mathfrak{M}) называется *обобщенным в узком смысле ПНЗ*, если существует целое число $r \geq 2$ такое, что для всех $M \in \mathfrak{M}$

$$\mu(\xi_{t+1} \in M | x^t, y^t) \equiv \mu(\xi_{t+1} \in M | y_{t-r}^t), \quad t \geq r.$$

Обобщенные в узком смысле ПНЗ обладают некоторым свойством однородности: повторение одного и того же действия y делает распределение $\mu(\cdot | yy \dots y)$ неизменным во времени. Примерами таких процессов служат функционалы на ОПНЗ, имеющие вид

$$\varphi_t = \varphi(\xi_t, y_{t-1}, \dots, y_{t-r}).$$

В качестве других примеров укажем процессы, задаваемые уравнением

$$\xi_t = g(y_{t-1}, \dots, y_{t-r}; \zeta_t),$$

где ζ_t — последовательность независимых одинаково распределенных случайных величин. Практически интересны скалярные процессы, задаваемые соотношениями

$$\xi_t = a_1 y_{t-1} + \dots + a_r y_{t-r} + \zeta_t.$$

К дальнейшему расширению класса управляемых процессов приходим, рассматривая функционалы на ОПНЗ

$$\varphi_t = \varphi(\xi_t; y_{t-1}, y_{t-2}, \dots, y_1, y_0) = \varphi(\xi_t; y^{t-1}).$$

Подобного вида процессы не обладают, вообще говоря, свойством однородности во времени.

Процесс ξ_t в фазовом пространстве (X, \mathfrak{M}) называется *обобщенным* в широком смысле ПНЗ, если для всех $M \in \mathfrak{M}$

$$\mu(\xi_{t+1} \in M | x^t, y^t) \equiv \mu(\xi_{t+1} \in M | y^t), \quad t \geq 0.$$

Легко видеть, что функционалы $\phi = \phi(\xi_t)$ на обобщенных в том или ином смысле ПНЗ являются обобщенными в том же смысле ПНЗ, что и исходный процесс ξ_t . Поэтому далее будем считать X принадлежащим числовой прямой.

Подобно тому, как при управлении конкретными ОПНЗ характерные цели обеспечиваются с помощью программных стратегий (притом стационарных, т. е. заключающихся в повторении одного и того же действия), так и для обобщенных ПНЗ обычно достаточно рассматривать лишь программные стратегии. Ограничивааясь таковыми, введем порождаемые ими меры и соответственно математическое ожидание. Математическое ожидание функционала $\varphi_t = \varphi(\xi_t)$ на процессах введенных двух типов имеет вид

$$E\varphi_{t+1} = W(y_{t-r}^t)$$

для обобщенных в узком смысле ПНЗ и

$$E\varphi_{t+1} = W(y^t)$$

для обобщенных в широком смысле ПНЗ (мы назовем их *средними выигрышами*). К этим функциям отнесем цель управления. Рассматривая обобщенные в узком смысле ПНЗ, попробуем сначала максимизировать средний выигрыш $W(y_{t-1}, \dots, y_{t-r})$, т. е. отыскать набор действий $y = \{y^{(1)}, \dots, y^{(r)}\}$, $y^i \in Y$, который примененный последовательно: $y_{t-r} = y^{(1)}$, $y_{t-r+1} = y^{(2)}$, ..., $y_{t-1} = y^{(r)}$, доставит максимум этой функции. Легко понять, что эта цель может приводить к большим суммарным потерям за длительный период управления. В самом деле, пусть набор действий y в момент времени t обеспечивает наибольший средний выигрыш, но в следующий момент времени $t + 1$ вместо этого набора появится новый $y' = \{y^{(2)}, \dots, y^{(r)}, y^{(r+1)}\}$, в результате действия которого

средний выигрыш может оказаться минимальным. Приведем несколько примеров.

Пусть множество $Y = \{y_1, \dots, y_k\}$ конечно и процесс ξ_t детерминирован: $\xi_t = y_{t-1} - y_{t-2}^2$ (т. е. $r = 2$). Предположим, что значения управлений $y' = 0$ и $y'' = 2$ ($k = 2$). Тогда средний выигрыш $W_t = y_{t-1} - y_{t-2}^2$ максимален и равен 2 при наборе $y = \{y'', y'\}$. В следующий момент времени $W(t+1) = y_t - y_{t-1}^2$, и если $y_t = 0$, то $W(t+1) = -4$ и весь доход на предыдущем шаге аннулирован. С течением времени при чередовании управлений y' , y'' суммарный доход будет стремиться к $-\infty$. Поэтому оказывается выгоднее воспользоваться программной стратегией $y_t \equiv y'$. При другом множестве управлений $Y = \{y^{(1)} = 1, y^{(2)} = 1/2, y^{(3)} = 0\}$ максимум среднего выигрыша, равный 1, достигается при наборе $\{y^{(3)}, y^{(1)}\}$, но суммарный выигрыш при его повторениях равен на интервалах четной длины 1, а нечетной — 0. Набор $\{y^{(2)}, y^{(1)}\}$ обеспечивает средний выигрыш $3/4$, при неограниченном его повторении суммарный средний выигрыш растет линейно. Итак, оказывается, что цель — максимизация среднего выигрыша — здесь неестественна.

Управление обобщенным (в узком смысле) ПНЗ представляет собой развивающийся во времени процесс. Совершенные к моменту t действия y_{t-1}, \dots, y_{t-r} в количестве r сменяются к следующему моменту $t + 1$ на иные действия $y_t, y_{t-1}, \dots, y_{t-r+1}$. Последствия воздействия нового набора могут резко отличаться от последствий предыдущего набора. Исключением служит выбор во все моменты времени одного и того же действия, но такие «удобные» наборы могут оказаться самыми невыгодными с точки зрения поставленной цели управления.

Мы будем пользоваться иногда таким обозначением обобщенных в узком смысле ПНЗ:

$$\xi_t(y_{t-r-1}^{t-1}) = \xi_t(y_{t-1}, \dots, y_{t-r}),$$

явно показывающим приложенные за предыдущие r тактов времени действия.

Перед формулированием цели управления обобщенными ПНЗ введем новое понятие. Зададим целое число $l > r$.

Циклическим выигрышем (глубины l за набор $\mathbf{y} = (y_{i_1}, \dots, \dots, y_{i_l})$) назовем функционал

$$\begin{aligned}\varphi^{(l)}(\mathbf{y}) = & \frac{1}{l} [\xi_t(y_{i_r}, \dots, y_{i_1}) + \xi_{t+1}(y_{i_{r+1}}, \dots, y_{i_2}) + \dots \\ & \dots + \xi_{t+l-r}(y_{i_l}, y_{i_{l-1}}, \dots, y_{i_{l-r+1}}) + \\ & + \xi_{t+l-r+1}(y_{i_1}, y_{i_2}, \dots, y_{i_{l-r+2}}) + \dots \\ & \dots + \xi_{t+l-1}(y_{i_{r-1}}, \dots, y_{i_1}, y_{i_l})],\end{aligned}$$

который получается в результате применения управлений $y_{i_1}, y_{i_2}, \dots, y_{i_l}, y_{i_1}, \dots, y_{i_{r-1}}$ слева направо одно за другим. Очевидно, наборы $\mathbf{y} = (y_{i_1}, \dots, y_{i_l})$ и $\mathbf{y}' = (y_{i_m}, \dots, y_{i_{l-1}}, y_{i_l}, y_{i_1}, \dots, y_{i_{m-1}})$ приводят к одному и тому же значению циклического выигрыша.

Средним циклическим выигрышем назовем математическое ожидание

$$\begin{aligned}E\varphi^{(l)}(\mathbf{y}) = & \frac{1}{l} [E\xi_t(y_{i_r}, \dots, y_{i_1}) + \dots \\ & \dots + E\xi_{t+l-1}(y_{i_{r-1}}, \dots, y_{i_1}, y_{i_l})].\end{aligned}$$

Зададим целью управления обобщенным ПНЗ максимизацию среднего циклического выигрыша. Такую цель можно понимать в двух смыслах: максимизация $E\varphi^{(l)}(\mathbf{y})$ либо при данном l , либо по всем l . Далее всегда имеется в виду второй из них.

Ниже будет показано, что адаптивная система, которая для класса обобщенных ПНЗ обеспечивает эту цель, приводит к максимизации с точностью до любого фиксированного $\varepsilon > 0$ величины

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t,$$

часто встречающейся в задачах оптимизации и управления. Максимизировать естественно в классе $\{\mathbf{y}^\infty\}$ программных стратегий. В пользу этого соглашения, кроме интуитивных соображений (связанных с тем, что вероятности $\mu(\cdot | \mathbf{y}^t)$ зависят лишь от прошлых управлений, но не от значений самого процесса), говорят результаты гл. IX (см. § 2).

§ 2. Обобщенные в узком смысле ПНЗ

В этом параграфе мы рассматриваем обобщенные в узком смысле процессы. Решается вопрос о существовании универсальной длины l_0 набора управлений (по которому строится циклический выигрыш), позволяющий максимизировать $E_{\varphi^{(l)}}(\mathbf{y})$ для всех l .

Предположим до конца главы, что пространство управлений конечно: $Y = \{y_1, \dots, y_k\}$.

Из средних выигрышей $W(y_{i_r}, \dots, y_{i_1}) = E\xi(y_{i_r}, \dots, y_{i_1})$ образуем «объемную» r -мерную матрицу порядка $k \times k \times \dots \times k$ и обозначим эту матрицу *средних значений* $M(k, r)$. Элемент, расположенный на «пересечении» i_1, i_2, \dots, i_r «строк», есть $W(y_{i_r}, \dots, y_{i_1})$.

Каждому набору управлений $\mathbf{y} = (y_{i_1}, \dots, y_{i_r})$ соответствует на матрице средних значений «путь» $\Pi(\mathbf{y})$, состоящий из тех элементов матрицы $M(k, r)$, которые отвечают r -м управлениям из набора \mathbf{y} , т. е. $(y_{i_r}, \dots, y_{i_1}), (y_{i_{r+1}}, \dots, y_{i_2})$ и т. д. Рассматривая набор \mathbf{y} как замкнутый (его последняя компонента y_{i_r} «примыкает» к первой y_{i_1}), после обхода его находим средний циклический выигрыш, суммируя пройденные элементы матрицы $M(k, r)$. Ясно, что между наборами \mathbf{y} и путями (замкнутыми) $\Pi(\mathbf{y})$ на матрице $M(k, r)$ существует взаимно однозначное соответствие. Скажем, что путь $\Pi(\mathbf{y})$ имеет самопересечение, если некоторый элемент матрицы встречается в нем дважды.

Теорема 1. *Максимум функции $E_{\varphi^{(l)}}(\mathbf{y})$ достигается на пути длины не более чем k^r .*

Доказательство. Для каждого $l \geq 1$ существует такой путь $\Pi(\mathbf{y}^0)$ длины l , на котором достигается максимум $E_{\varphi^{(l)}}(\mathbf{y})$. Пути длины, превосходящей k^r (количество элементов матрицы $M(k, r)$), непременно имеют самопересечения. Их можно отбросить. В самом деле, пусть путь $\Pi(\mathbf{y})$ длины l имеет одно самопересечение, причем l_1 и l_2 ($l_1 + l_2 = l$) — длины участков этого пути \mathbf{y}_1 и \mathbf{y}_2 без самопересечения. Тогда справедливо легко доказываемое неравенство

$$E_{\varphi^l}(\mathbf{y}) \leq \max(E_{\varphi^{(l_1)}}(\mathbf{y}_1), E_{\varphi^{(l_2)}}(\mathbf{y}_2)).$$

Отсюда непосредственно следует, что средний циклический выигрыш по пути на матрице $M(k, r)$ с одним самопересечением мажорируется средним циклическим выигрышем по пути без самопересечения меньшей длины. Сказанное непосредственно переносится на пути с несколькими самопересечениями. Доказательство теоремы завершается указанием, что длина пути без самопересечения на матрице $M(k, r)$ не превосходит k^r .

Мы хотим теперь найти, если это возможно, такие длины пути $l(k, r)$, чтобы для всех обобщенных в узком смысле ПНЗ с одними и теми же значениями k и r средние циклические выигрыши достигали максимума на каком-нибудь пути длины $l(k, r)$. Через $l_0(k, r)$ обозначим наименьшую среди таких длин $l(k, r)$.

Под записью $\max_{\mathbf{y}} E_{\varphi^{(l)}}(\mathbf{y})$ понимается максимизация по наборам фиксированной длины l , а запись $\max_l E_{\varphi^{(l)}}(\mathbf{y})$ означает, что максимизация производится по наборам произвольной длины.

Лемма. *Если* $\max_l E_{\varphi^{(l)}}(\mathbf{y})$ *достигается на пути* $\Pi(\mathbf{y}_0)$ *длиной* l_0 , *то при любом целом* $n \geq 1$ *справедливо равенство*

$$E_{\varphi^{(l_0)}}(\mathbf{y}_0) = \max E_{\varphi^{(nl_0)}}(\mathbf{y}),$$

где $\mathbf{y} = ny_0$ есть n -кратное повторение набора y_0, y_0, \dots, y_0 .

Доказательство. В силу равенства $E_{\varphi^{(l_0)}}(\mathbf{y}_0) = \max_l E_{\varphi^{(l)}}(\mathbf{y})$ имеем для любого набора \mathbf{y} длины nl_0

$$E_{\varphi^{(l_0)}}(\mathbf{y}_0) \geq E_{\varphi^{(nl_0)}}(\mathbf{y}),$$

для n -кратно повторенного набора \mathbf{y}_0 , обозначенного через ny_0 ,

$$E_{\varphi^{(nl_0)}}(ny_0) = E_{\varphi^{(l_0)}}(\mathbf{y}_0).$$

Отсюда непосредственно получаем утверждение леммы.

Через p_1, \dots, p_m обозначим все простые числа, не превосходящие k^r . Выберем числа n_1, \dots, n_m так, чтобы выполнялись условия $p_i^{n_i} \leq k^r < p_i^{n_i+1}$, $i = 1, \dots, m$. Положим

$$l_0(k, r) = \prod_{i=1}^m p_i^{n_i}.$$

Легко видеть, что $l_0(k, r)$ — наименьшее число, кратное всем числам, не превосходящее k^r .

Введем класс $Q(X; k, r)$ обобщенных в узком смысле ПНЗ с одним и тем же фазовым пространством X , пространством управлений из k элементов $Y = \{y_1, \dots, y_k\}$ и вероятностями $\mu(\cdot | y_{t-r}^t)$, существенно зависящими от r предыдущих управлений.

Теорема 2. Для любого процесса из класса $Q(X; k, r)$ найдется хоть один набор управлений y длины $l_0(k, r)$ такой, что на нем достигается $\max_l E_{\varphi^{(l)}}(y)$. Не существует меньшей, чем $l_0(k, r)$, длины с указанным свойством.

Доказательство. Согласно теореме 1 для каждого процесса из $Q(X; k, r)$ существует набор длины $l_1 \leq k^r$, на котором достигается $\max_l E_{\varphi^{(l)}}(y) = \max_l E_{\varphi^{(l_1)}}(y)$. Мы знаем, что $l_0(k, r)$ кратно l_1 . Следовательно, по лемме

$$\max_l E_{\varphi^{(l_1)}}(y) = \max_l E_{\varphi^{(l_0(k, r))}}(y).$$

Пусть $l_2 < l_0(k, r)$ и $l < k^r$ такое, что l_2 не кратно l . Ясно, что существует процесс $\xi_t(\omega) \in Q(X; k, r)$ такой, что минимальная длина набора, на котором достигается $\max_l E_{\varphi^{(l)}}(y)$, равна l и на наборах длины, не кратной l , $\max_l E_{\varphi^{(l)}}(y)$ не достигается. Этим доказывается второе утверждение теоремы. Теорема доказана.

Ниже нам потребуется одно вспомогательное неравенство, справедливое для процессов с неотрицательными значениями (т. е. $X \subseteq [0, \infty]$).

Пусть $\tilde{\Pi}$ — путь длины l , вообще говоря, незамкнутый, на матрице $M(k, r)$, проходящий только через элементы

$$(y_{i_r}, \dots, y_{i_1}), (y_{i_{r+1}}, \dots, y_{i_2}), \dots, (y_{i_{l+r-1}}, \dots, y_{i_l}).$$

Их совокупность обозначим \tilde{z}' . Добавлением $r - 1$ элементов

$$\begin{aligned} \tilde{z}'' = [(y_{i_1}, y_{i_{l+r-1}}, \dots, y_{i_{l+1}}), (y_{i_2}, y_{i_1}, y_{i_{l+r-1}}, \dots, y_{i_{l+2}}), \dots \\ \dots, (y_{i_{r-1}}, y_{i_{r-2}}, \dots, y_{i_1}, y_{i_{l+r-1}})] \end{aligned}$$

превращаем путь $\tilde{\Pi}$ в замкнутый $\Pi(z)$, отвечающий набору $z = \{z', z''\}$. Перенумеруем элементы пути и через W_j

обозначим значение функции $W(y^{(1)}, \dots, y^{(r)})$ на j -м элементе. Согласно теореме 2

$$\frac{1}{l+r-1} \sum_{j \in \Pi(z)} W_j \leq \max E_{\varphi^{(l_0)}}(y).$$

Левая часть этого неравенства равна

$$\frac{1}{l+r-1} \sum_{j \in \Pi(z)} W_j = \frac{1}{l+r-1} \left(\sum_{j \in \bar{\Pi}} W_j + \sum_{j \in \Pi(z) \setminus \bar{\Pi}} W_j \right).$$

Поэтому

$$\sum_{j \in \bar{\Pi}} W_j + \sum_{j \in \Pi(z) \setminus \bar{\Pi}} W_j \leq (l+r-1) \max E_{\varphi^{(l_0)}}(y).$$

Мы усилим неравенство (в силу неотрицательности процесса и средних выигрышей), если отбросим в левой части вторую сумму. Интересующее нас неравенство имеет вид

$$\frac{1}{l} \sum_{j \in \bar{\Pi}} W_j \leq \left(1 + \frac{r-l}{l}\right) \max E_{\varphi^{(l_0)}}(y). \quad (1)$$

§ 3. Обобщенные в широком смысле ПНЗ

Будем далее часто обозначать процессы ξ_t , описываемые вероятностями $\mu(\cdot | y^t)$, через $\xi_t(y^{t-1})$, явным образом указывая совершенные до момента времени t управления, которые задают вероятностные свойства таких процессов. Рассматриваются классы скалярных процессов этого типа. Избегая тривиальных случаев, примем, что их распределения ни при каком t не сосредоточены в крайних точках $x = 0$ и $x = A$ множества значений $[0, A] = X$. Ниже это допущение специально оговариваться не будет. Мы выдвигаем в качестве цели управления максимизацию

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E \xi_t.$$

Точнее говоря, задача состоит в синтезе семейства адаптивных систем, которые обеспечивают значения величины

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E \xi_t,$$

отличающиеся не более чем

на заданное ε от $\sup_{(y^\infty)} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E \xi_t$, причем здесь рассматриваются лишь программные стратегии $\sigma = (y^\infty)$. Соображения в пользу ограничения такими стратегиями были высказаны в § 1.

Средние выигрыши являются математическими ожиданиями, вычисленными относительно меры, которая порождена программной стратегией. Пусть $y^{t-1} = (y_{t-1}, \dots, y_1, y_0)$ — t -мерный набор управлений, совершенных в моменты $t = 1, \dots, 1, 0$. Условимся обозначать

$$W(y^{t-1}) = E(\xi_t | y^{t-1}) = \int_X x \mu(dx | y^{t-1}).$$

Тогда цель управления заключается в максимизации $\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T W(y^t)$ (с точностью до заданного ε) функции $W(y^t)$.

Укажем пример процесса, для которого никакая адаптивная система не может гарантировать достижение этой цели. Пусть управление y принимает два значения 0 и 1. Примем, что средний выигрыш $W(y^t)$, где $y^t = (y_{i_t}, y_{i_{t-1}}, \dots, y_{i_1}, y_{i_0})$ при всех t равен

$$W(y^t) = 0, \quad y_{i_0} y_{i_1} \dots y_{i_{t-1}} y_{i_t},$$

который записан в двоичной системе. Ясно, что адаптивная система, выбирая оба возможных действия $y = 0$ и $y = 1$ (по крайней мере на начальном этапе функционирования), не с состояниями обеспечить цель управления. Поэтому следует наложить ограничения на класс управляемых процессов, суть их заключается в требовании угасания с течением времени влияния ранних управлений (в гл. II мы уже допустили, что это условие необходимо для существования адаптивной системы).

Условие на класс обобщенных в широком смысле ПНЗ заключается в следующем: при любых целых положительных t, t' и выполнении равенства $y^t = y^{t+t'}$ имеем

$$|W(y^{t+t'}) - W(y^t)| < \alpha(t), \quad (1)$$

где $\lim_{t \rightarrow \infty} \alpha(t) = 0$, причем стремление к 0 монотонное.

Обобщенные в широком смысле ПНЗ, подчиненные этому условию, назовем *обобщенными ПНЗ*.

Очевидно, что обобщенные в узком смысле ПНЗ являются обобщенными ПНЗ. Задача синтеза для них адаптивных систем есть частный случай задачи синтеза адаптивных систем для обобщенных ПНЗ.

Ограничиваюсь по-прежнему конечными $Y = \{y_1, \dots, y_k\}$, установим некоторые общие свойства обобщенных ПНЗ. Основную роль играет понятие, обобщающее соответствующее понятие из § 2.

Средним циклическим выигрышем (порядка r за набор y_i длины l) называется сумма средних выигрышей*, зависящих от r аргументов

$$f(y_l, r) = \frac{1}{l} \sum_{j \in \Pi(y)} W_j,$$

взятых вдоль замкнутого пути $\Pi(y)$ на матрице $M(k, r)$.

В основе дальнейших результатов лежит такая

Л е м м а. *Существует предел*

$$\lim_{r \rightarrow \infty} \max f(y_{l_0(k, r)}, r) = f^*.$$

Доказательство. Зафиксируем $\epsilon > 0$ и через $r_1 = r_1(\epsilon)$ обозначим наименьшее целое число такое, что $\alpha(r_1) < \epsilon/2$. Пусть $r_2 > r_1$. Сопоставим порядкам наборов r_1 и r_2 соответственно длины наборов $l_1 = l_0(k, r_1)$, $l_2 = l_0(k, r_2)$, и пусть y_i , $i = 1, 2, \dots$ — наборы длин $l_0(k, r_i)$, на которых достигаются максимумы $f(y_{l_0(k, r_i)}, r_i)$. По определению обобщенного ПНЗ и выбору r_1, r_2 имеем

$$E(\xi_{r_1} | y_{i_{l_2}}, y_{i_{l_2-1}}, \dots, y_{i_{l_2-r_1+1}}) - E(\xi_{r_2} | y_{i_{l_2}}, \dots, y_{i_{l_2-r_2+1}}) < \frac{\epsilon}{2}.$$

Такие неравенства верны для каждой пары взаимно соответствующих слагаемых в $f(y_{l_1}, r_1)$ и $f(y_{l_2}, r_2)$. Поэтому можно утверждать, что

$$f(y_2, r_2) \leq f(y_2, r_1) + \frac{\epsilon}{2} \leq \max f(y_{l_1}, r_1) + \frac{\epsilon}{2},$$

$$f(y_1, r_1) \leq f(y_1, r_2) + \frac{\epsilon}{2} \leq \max f(y_{l_2}, r_2) + \frac{\epsilon}{2}.$$

*) То есть математические ожидания $E(\xi_{r+1} | y_{i_1}, \dots, y_{i_r})$.

Из них выводим при $r_2 > r_1(\varepsilon)$

$$|\max f(\mathbf{y}_{t_1}, r_1) - \max f(\mathbf{y}_{t_2}, r_2)| < \varepsilon,$$

т. е. интересующий нас предел действительно существует.

Теорема. Для любого обобщенного ПНЗ с неотрицательными значениями $\xi(y^{t-1})$ справедливо равенство

$$\sup_{y^\infty} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t(y^{t-1}) = f^*.$$

Доказательство состоит из двух этапов.

1. Зададим $\varepsilon > 0$. Найдется бесконечная последовательность управлений $y^\infty(\varepsilon)$ такая, что

$$\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t(y^{t-1}(\varepsilon)) + \frac{\varepsilon}{2} > \sup_{y^\infty} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t(y^{t-1}), \quad (2)$$

где $y^{t-1}(\varepsilon)$ — начальный отрезок длины t последовательности y_ε^∞ . По определению обобщенного ПНЗ существует $r' = r'(\varepsilon)$, для которого

$$E\xi_t(y^{t-1}_\varepsilon) \leq E\xi_{r+1}(y^{(\varepsilon)}_{t-1}, \dots, y^{(\varepsilon)}_{t-r}) + \frac{\varepsilon}{8}.$$

Условимся по-прежнему писать $E(\xi_{r+1} | y^{t-1}_{t-r-1}) = W(y^{t-1}_{t-r})$ при $t > r$. Тогда

$$\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t(y^{t-1}(\varepsilon)) \leq \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=r+1}^T W(y^{t-1}_{t-r}(\varepsilon)) + \frac{\varepsilon}{8}. \quad (3)$$

Последовательность функций $W(y^{t-1}_{t-r}(\varepsilon))$ будем трактовать как детерминированный обобщенный в узком смысле ПНЗ, зависящий в каждый момент от r' предыдущих управлений.

Тогда сумму $\sum_{t=1}^T W(y^{t-1}_{t-r}(\varepsilon))$ представляем себе взятой «вдоль» бесконечного пути (на матрице $M(k, r)$), отвечающего последовательности $y^\infty(\varepsilon)$. Значит, при каждом T справедливо неравенство (1), доказанное в конце § 2, которое в наших обозначениях принимает вид

$$\frac{1}{T} \sum_{t=1}^T W(y^{t-1}_{t-r}(\varepsilon)) \leq \left(1 + \frac{r' - 1}{T}\right) \max f(\mathbf{y}_{t_0(k, r')}, r').$$

Переходя здесь к верхнему пределу (при $T \rightarrow \infty$) и принимая во внимание неравенство

$$\left| \lim_{r \rightarrow \infty} \max f(\mathbf{y}_{l_0}, r) - \max f(\mathbf{y}_{l_0(k, r')}, r') \right| < \frac{\varepsilon}{4},$$

следующее из сделанного выбора r' , получаем

$$\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T W(y_{t-r'}^{t-1}) \leq \lim_{r \rightarrow \infty} \max f(\mathbf{y}_{l_0}, r) + \frac{\varepsilon}{4}.$$

Сопоставим это неравенство с (2) и (3). В силу произвольности выбора ε приходим к желаемой оценке

$$\sup_{y^\infty} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E(\xi_t(y^{t-1})) \leq \lim_{r \rightarrow \infty} \max f(\mathbf{y}_{l_0(k, r)}, r).$$

2. Будем доказывать противоположное неравенство, что приведет нас к утверждению теоремы. Снова зафиксируем $\varepsilon > 0$ и выберем r' так, чтобы

$$\max f(\mathbf{y}_{l_0(k, r')}, r') + \varepsilon > \lim_{r \rightarrow \infty} \max f(\mathbf{y}_{l_0(k, r)}, r)$$

и, кроме того, $\alpha(r') < \varepsilon/2$. Образуем бесконечную последовательность управлений $y^\infty(\varepsilon)$ повторением набора \mathbf{y}^0 длины $l_0(k, r')$, который доставляет максимум среднему циклическому выигрышу $f(\mathbf{y}_{l_0(k, r')}, r')$. Из определения (1) обобщенного ПНЗ имеем

$$\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t(y^{t-1}(\varepsilon)) \geq \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T W(y_{t-r+1}^t(\varepsilon)) - \frac{\varepsilon}{2}.$$

Принимая во внимание

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T W(y_{t-r+1}^t(\varepsilon)) = f(\mathbf{y}^0, r') = \max f(\mathbf{y}_{l_0(k, r')}, r'),$$

получаем из предыдущего неравенства

$$\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t(y^{t-1}(\varepsilon)) \geq \lim_{r \rightarrow \infty} \max f(\mathbf{y}_{l_0(k, r')}, r) - \frac{3}{2}\varepsilon,$$

а в силу произвольности ϵ

$$\sup_{y \in \mathcal{Y}} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E \xi_t(y^{t-1}) \geq \lim_{r \rightarrow \infty} \max f(\mathbf{y}, r).$$

Этим завершено доказательство теоремы.

В дальнейшем изложении будем употреблять для обобщенных ПНЗ термин «циклический выигрыш», охватывающий введенное в § 1 одноименное понятие. Смысл его таков: пусть с момента $t+1$ управление процессом осуществляется действиями из набора $\mathbf{y}_l = \{y_{i_1}, \dots, y_{i_l}\}$. Циклический выигрыш равен сумме l значений процесса, начиная с момента $t+r+1$.

Пусть задано $\epsilon > 0$. Согласно определению обобщенного ПНЗ (см. (1)) найдется такое значение t' , что при любом t и любых управлениях таких, что $y_t^{t+t'} = y^{t'}$, окажется справедливым неравенство

$$|E(\xi_{t+t'+1} | y^{t+t'}) - E(\xi_{t'+1} | y^{t'})| \leq \alpha(t') < \epsilon.$$

Параметр $r = r(\epsilon)$ выберем наименьшим среди t' , удовлетворяющих этому неравенству. Теперь дадим формальное определение.

Циклический выигрыш задается выражением

$$\varphi_t(\mathbf{y}_l, r) = \frac{1}{l} \sum_{j=1}^l \xi_{t+r+j}(\tilde{y}_j(\mathbf{y}_l, r, t)),$$

где \tilde{y}_j — последовательность управлений длины $t+r+j$ — имеет следующий вид. Ее первые t элементов произвольны, а за ними следуют действия набора $\mathbf{y}_l = (y_{i_1}, \dots, y_{i_l})$ в таком порядке:

$$y_{i_{l-r+1}}, \dots, y_{i_l}, y_{i_1}, \dots, y_{i_j}.$$

Связь между циклическим выигрышем $\varphi_t(\mathbf{y}_l, r)$ и средним циклическим выигрышем $f(\mathbf{y}_l, r)$ при одинаковых значениях параметров (r, l) и наборе \mathbf{y}_l указывает такое соотношение: при всех t , каков бы ни был обобщенный ПНЗ,

$$\lim_{r \rightarrow \infty} [E \varphi_t(\mathbf{y}_l, r) - f(\mathbf{y}_l, r)] = 0,$$

где $l_0 = l_0(k, r)$. Это соотношение непосредственно следует из определения рассматриваемого класса процессов.

§ 4. Синтез автоматов для управления обобщенными ПНЗ

Пусть ξ_t — обобщенный ПНЗ с фазовым пространством $X = [0, A]$ — конечным отрезком числовой прямой и пространством управлений $Y = (y_1, \dots, y_k)$. Наложенные условия гарантируют существование всех вводимых математических ожиданий. Через $\varphi_t(y_i, r)$ снова обозначим циклический выигрыш $\varphi_t(y_i, r)$. Его значение φ называется *приемлемым*, если независимая от ξ_t случайная величина ς , равномерно распределенная на отрезке $[0, 1]$, приняла значение из полуинтервала $[0, \varphi/A]$, и *неприемлемым*, если $\varsigma \geq \varphi/A$. Смысл этого определения заключается в том, что с вероятностью φ/A вырабатывается поощрение, а с дополнительной — наказание.

Укажем конструкцию семейства адаптивных систем, которые обеспечивают цель: для любого $\varepsilon > 0$ максимизировать с точностью до ε величину

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t,$$

каков бы ни был обобщенный ПНЗ ξ_t из класса $Q([0, A]; k)$ таких процессов с $X = [0, A]$ и $Y = (y_1, \dots, y_k)$. Это семейство систем (« ε -оптимальное семейство») строится из модифицированных автоматов $\tilde{A}_{k,n}$, которые были определены в § 4 гл. III, увеличением количества состояний и усложнения каждого из них. Изложение упростится, если выбрать какое-либо конкретное семейство модифицированных автоматов. В основу дальнейших рассуждений положим автоматы $\tilde{D}_{k,n}$.

Синтезируются автоматы $(CD)_{k,n}^{(l,r)}$. Их входными сигналами являются значения процесса ξ_t — числа из отрезка $[0, A]$. Выходные сигналы — управление $y \in Y$. Остается задать множество состояний $S_{n,k}^{(l)}$, а с ним функции перехода и выхода. Параметр n означает, как и в гл. III, глубину памяти автомата. Множество $S_{n,k}^{(l)}$ состоит из центрального состояния s_0 , оно же начальное, и k^l групп

периферических состояний S_{i_1}, \dots, i_l , содержащих каждая по n^l элементов

$$s = (d; y_{i_1}, \dots, y_{i_l}; m), \quad d = 1, \dots, l; \quad m = 1, \dots, n.$$

В центральном состоянии s_0 с одинаковыми вероятностями выбирается один из наборов управлений $y_i = (y_{i_1}, \dots, y_{i_l})$ длины l . Действия этого набора совершаются в таком порядке:

$$y_{i_{l-r+1}}, \dots, y_{i_l}, y_{i_1}, \dots, y_{i_l}, \quad l \geq r.$$

По последним l значениям процесса ξ_t вычисляется циклический выигрыш $\varphi_t(y_i, r)$. Разумеется, значения процесса в каждый момент времени зависят от всех прилагавшихся ранее управлений. Если полученный циклический выигрыш оказался неприемлемым, автомат остается в состоянии s_0 , в котором избирается новый набор управлений. Если же циклический выигрыш приемлем, автомат переходит в группу состояний S_{i_1}, \dots, i_l и оказывается во входном состоянии $(1; y_{i_1}, \dots, y_{i_l}; 1)$. В состоянии $(d; y_{i_1}, \dots, y_{i_l}; m)$, $1 \leq m \leq n$, совершается действие y_{i_d} и при $d \neq l$ автомат переходит из этого состояния в $(d+1; y_{i_1}, \dots, y_{i_l}; m)$. В случае $d = l$ совершается y_{i_l} и по последним l значениям процесса вычисляется циклический выигрыш. Если он приемлем, автомат переходит в «самое глубокое» состояние $(1; y_{i_1}, \dots, y_{i_l}; n)$, а в противном случае — в $(1; y_{i_1}, \dots, y_{i_l}; m-1)$, причем условимся считать, что $(1; \dots; 0)$ является центральным состоянием s_0 .

Каждый автомат $(CD)_{k,n}^{(l,r)}$ реализует стратегию для класса $Q([0, A]; k)$ обобщенных ПНЗ. Можно вычислить математическое ожидание $E\xi_t$ по мере, порожденной этой стратегией. Нас далее интересует предельное математическое ожидание циклического среднего выигрыша $\lim_{l \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l E\xi_{t+i}$. Однако этот предел, вообще говоря, не существует для процессов из $Q([0, A]; k)$. Поэтому мы

вводим величину

$$W(n; r, l) = \lim_{t \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l E\xi_{t+i},$$

где n — глубина памяти автомата $(CD)_{k, n}^{(l, r)}$.

Лемма. Для всех обобщенных ПНЗ из класса $Q([0, A]; k)$

$$W(n; r, l) \leq \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t.$$

Если существует $\lim_{t \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l E\xi_{t+i}$, то

$$W(n; r, l) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t.$$

Доказательство. Образуем величину $e_t = \sum_{i=1}^l E\xi_{t+i}$.

Имеем

$$\frac{1}{T} \sum_{t=1}^T E\xi_t = \frac{1}{T} \sum_{i=0}^{[T/l]} e_{il} + \frac{1}{T} \sum_{i=[T/l]+1}^T E\xi_i. \quad (1)$$

Вторая сумма справа стремится к нулю когда $T \rightarrow \infty$ (в ней не более, чем l ограниченных слагаемых). Зафиксируем $\varepsilon > 0$ и найдем такое t_ε , что $e_t > lW(n; r, l) - \varepsilon$ при всех $t > t_\varepsilon$. Тогда первая сумма в правой части принимает вид

$$\frac{1}{T} \sum_{i=0}^{[T/l]} e_{il} = \frac{1}{T} \sum_{i=0}^{[T_\varepsilon/l]} e_{il} + \frac{1}{T} \sum_{i=[T_\varepsilon/l]+1}^{[T/l]} e_{il},$$

где $T_\varepsilon = lt_\varepsilon$. Первая сумма стремится к нулю, а вторая сумма не менее $W(n; r, l) - \varepsilon$. Произвольность ε влечет справедливость первого утверждения теоремы. Второе утверждение вытекает, согласно (1), из того, что $e_t \rightarrow lW(n; r, l)$, ибо

$$\frac{1}{T} \sum_{i=0}^{[T/l]} e_{il}$$

есть чезаровское среднее сходящейся последовательности.

§ 5. Свойство адаптивности автоматов $(CD)_{k,n}^{(l,r)}$

Мы покажем, что автоматы $(CD)_{k,n}^{(l,r)}$ образуют в классе обобщенных ПНЗ $Q([0, A]; k)$ ϵ -оптимальное семейство, т. е. с точностью до ϵ максимизируют средний циклический выигрыш. В качестве первого шага установим вспомогательный результат о цели управления, достигаемой модифицированными автоматами $\tilde{D}_{k,n}$ в классе обобщенных в широком смысле ПНЗ, подчиненных следующему условию: при некотором $\alpha > 0$, любом $y \in Y$ и любых t_1, t_2

$$|\mathbf{E}(\xi_{t_1} | y_{t_1-1} = y) - \mathbf{E}(\xi_{t_2} | y_{t_2-1} = y)| \leq \alpha. \quad (1)$$

Для ограниченных процессов ($0 \leq \xi_t \leq A$) это условие выполнено при $\alpha = A$. Ниже подразумевается, что α мало по сравнению с A . Примером таких процессов служит последовательность значений циклического выигрыша $\varphi_t(y_t, r)$ в моменты времени, кратные l . Управлением служит вектор-набор $y = (y_{i_1}, \dots, y_{i_l})$.

В интересах изложения мы рассмотрим здесь более широкий класс процессов. Они характеризуются условными распределениями общего вида $\mu(\cdot | x^t, y^t)$, принимают значения из отрезка $[0, A]$, имеют пространство управлений $Y = \{y_1, \dots, y_k\}$ и подчиняются условию (1).

Такие процессы назовем П-процессами.

Условимся еще обозначать

$$a_i = \mathbf{E}(\xi_i | y_i), \quad \bar{a} = \max_i a_i.$$

Пусть $\bar{a} < 1$ и $A = 1$. Поручим управление П-процессом ξ_t автомatu $\tilde{D}_{k,n}$.

Нас интересует оценка среднего времени $T_n(y)$ совершения действия y подряд. Обозначим через $v(x_t) = (v_1, \dots, v_m)$ последовательность поощрений и наказаний, вырабатываемую автоматом из входных сигналов $\xi_{t+1}, \dots, \xi_{t+m}$, от начала совершения действия y до момента смены его. Вероятность поощрения обозначим $p(1 | x_t)$, а наказания $p(0 | x_t)$. Здесь через x_t обозначена предысто-

рия процесса до момента времени t . Вероятность последовательности v равна

$$P(v) = \prod_{i=1}^m P(v_i | x_{t+i-1}).$$

Отметим три свойства таких последовательностей входных сигналов: 1) длина последовательности $m=1$ тогда и только тогда, когда $v_1=0$; 2) если $m>1$, то $m \geq n+1$ и справедливы равенства $v_{m-n}=1$, $v_{m-n+1}=\dots=v_{m-1}=v_m=0$; 3) среди элементов v_2, \dots, v_{m-n+1} есть не менее $\left[\frac{m-n-2}{n}\right]=q$ единиц (поощрений), отстоящих одна от другой не более чем на n позиций.

Введем $\{v_m(i_1, \dots, i_q)\}$ — множество входных последовательностей фиксированной длины $m>1$ таких, что в них на позициях i_1, i_2, \dots, i_q расположены единицы. Вероятность появления последовательности из этого множества равна, как легко убедиться,

$$\begin{aligned} P_\xi(\{v_m(i_1, \dots, i_q)\}) &= \\ &= p(1|x_t)p(1|x_{t+i_1})\dots p(1|x_{t+i_q})p(1|x_{t+m-n+1}) \times \\ &\quad \times p(0|x_{t+m-n+2})\dots p(0|x_{t+m}). \end{aligned}$$

Пусть теперь автомат $\tilde{D}_{k,n}$ управляет ОПНЗ η_t с тем же фазовым пространством, что у ξ_t , и с вероятностью $a+\alpha$, где $a=E(\xi_1|y)$, появления единицы на входе автомата при действии y , причем выполнены условия

$$1 > a + \alpha > a + \varepsilon \geq p(1|x_t) \geq a - \varepsilon > 0$$

при всех предысториях x_t . Здесь α и ε — фиксированные числа.

Нетрудно доказать, что для процесса η_t вероятность появления последовательности из множества $v_m(i_1, \dots, i_q)$ равна

$$P_{\eta_t}(\{v_m(i_1, \dots, i_q)\}) = (a + \alpha)^{\left[\frac{m-n-2}{n}\right] + 2} (1 - a - \alpha)^n.$$

Сравним вероятности $P_\xi(\cdot)$ и $P_\eta(\cdot)$. Во-первых, имеют место неравенства

$$\begin{aligned} \left(\frac{a-\varepsilon}{a+\alpha}\right)^{\left[\frac{m-n-2}{n}\right]+2} &\leqslant \\ &\leqslant \frac{p(1|x_t)p(1|x_{t+m-n+1})p(1|x_{t+i_1})\cdots p(1|x_{t+i_q})}{(a+\alpha)^{\left[\frac{m-n-2}{n}\right]+2}} \leqslant \\ &\leqslant \left(\frac{a+\varepsilon}{a+\alpha}\right)^{\left[\frac{m-n-2}{n}\right]+2}, \end{aligned} \quad (2)$$

правая часть стремится к 0 при $m \rightarrow \infty$ и фиксированном n . Кроме того, справедливы неравенства

$$\frac{1}{(1-a-\alpha)^n} \geqslant \frac{p(0|x_{t+m-n})\cdots p(0|x_{t+m})}{(1-a-\alpha)^n} \geqslant \left(\frac{1-a-\varepsilon}{1-a-\alpha}\right)^n > 1, \quad (3)$$

так как n фиксировано.

Следовательно, при всех m , начиная с некоторого m_0 , справедливо неравенство

$$\frac{P_\xi(\{v_m(i_1, \dots, i_q)\})}{P_\eta(\{v_m(i_1, \dots, i_q)\})} < 1, \quad m \geqslant m_0,$$

полученное из (2) и (3). В силу того, что левая часть в (2) монотонно стремится к 0, находим при $m \leqslant m_0$

$$\frac{P_\xi(\{v_m(i_1, \dots, i_q)\})}{P_\eta(\{v_m(i_1, \dots, i_q)\})} \geqslant 1.$$

Сопоставляя распределения вероятностей, порожденные П-процессом ξ_t и ОПНЗ η_t , мы заключаем, что математические ожидания $T_{x_t}^{(n)}$ и $T^{(n)}$ — средние времена до смены действий автоматом $D_{k,n}$, управляющим этими процессами, связаны неравенством

$$T_{x_t}^{(n)} \leqslant T^{(n)} = (1-a-\varepsilon)^{-n}.$$

Посредством аналогичных рассуждений получаем оценку снизу среднего времени

$$T_{x_t}^{(n)} \geqslant (1-a+\varepsilon)^{-n}.$$

Последние неравенства выводились в предположении, что фиксирована предыстория x_t . Проводя осреднение по всем возможным предысториям, получаем оценки средних времен до смены действия y_i : $(1 - a_i + \varepsilon)^{-n} \leq T^{(n)}(y_i) < (1 - a_i - \varepsilon)^{-n}$.

Нас будет интересовать величина $W(n; \xi_t; l, r)$, определяемая как нижний предел (по времени) математического ожидания выигрыша автомата $\tilde{D}_{k,n}$ при управлении Π -процессом ξ_t .

Лемма. *При управлении Π -процессом ξ_t с помощью автомата $\tilde{D}_{k,n}$ для любого $\varepsilon > 0$ найдется целое n_ε такое, что при всех $n > n_\varepsilon$ имеем*

$$W(n; \xi_t; l, r) \geq \bar{a} - 3\alpha - \varepsilon.$$

Доказательство. Пусть числа a_i пронумерованы в порядке убывания, $a_1 \geq a_2 \geq \dots \geq a_k$, и m — максимальный индекс такой, что $a_1 - \alpha \leq a_m + \alpha$. Тогда для средних времен совершения действий автомата $\tilde{D}_{k,n}$ имеем при $i > m$

$$T^{(n)}(i) \leq (1 - a_{m+1} - \alpha)^{-n},$$

а при $i \leq m$

$$T^{(n)}(i) \geq (1 - a_i + \alpha)^{-n}.$$

Введем теперь автомат $D_{k,n}$ (не модифицированный!), управляющий бинарным ОПНЗ ζ_t таким, что средние выигрыши за m первых действий равны $a_i - \alpha$, а за остальные $k - m$ действий равны $a_{r+1} + \alpha$. Вспомним, что (см. гл. III, § 2) средние времена совершения первых m действий этим автоматом равны соответственно $(1 - a_i + \alpha)^{-n}$, а остальных $(1 - a_{m+1} - \alpha)^{-n}$. Отсюда вытекает, что лучшие действия автомата $\tilde{D}_{k,n}$ совершают чаще автомата $D_{k,n}$, причем средний выигрыш за y_i , $i \leq m$, автомата $\tilde{D}_{k,n}$ не менее $a_i - \alpha$, а у $D_{k,n}$ он равен $a_i - \alpha$. Поэтому справедливы неравенства

$$W(n; \xi_t; l, r) \geq W(D_{k,n}; \zeta_t) > \bar{a} - 3\alpha - \varepsilon,$$

из которых вытекает утверждение леммы.

Отметим, что за числа a_i можно в лемме принять любые числа, которые при всех t удовлетворяют неравенствам $|a_i - E(\xi_t | y_i)| < \alpha$.

Вернемся к исследованию свойств автоматов $(CD)_{k,n}^{(l,r)}$, как управляющих систем для обобщенных ПНЗ из класса $Q([0, A]; k)$.

Теорема 1. *Пусть автомат $(CD)_{k,n}^{(l,r)}$ управляет обобщенным ПНЗ класса $Q([0, A]; k)$. Тогда для любого $\epsilon > 0$ найдутся такие n_ϵ и порядок r циклического выигрыша, что при всех $n > n_\epsilon$*

$$W(n; r, l_0(k, r)) > \max f(\mathbf{y}_{l_0(k, r)}, r) - \epsilon,$$

где $l_0(k, r)$ определено в теореме 2 из § 2.

Доказательство. Будем рассматривать автомат $(CD)_{k,n}^{(l,r)}$ только в моменты смены глубины состояния. Это означает переход к другому автомату $\bar{D}_{k,l,n}$, состояниями которого являются группы состояний $\{(d; y_{i_1}, \dots, y_{i_l}; m), d = 1, \dots, l\}$. Входными сигналами его служат циклические выигрыши, а точнее — «поощрение», если циклический выигрыш приемлем, и «наказание» в противном случае. Подобное толкование входных сигналов влечет за собою то, что автоматы $\bar{D}_{k,l,n}$ становятся изоморфными автомату $\bar{D}_{k,n}$, «действиями» которого служат наборы $y = (y_{i_1}, \dots, y_{i_l})$. Таким образом, $\bar{D}_{k,n}$ реализует стратегию управления П-процессом φ_t (с шагом l по времени) и допустимо воспользоваться леммой. Значит, для каждого ϵ' найдется такое $n_{\epsilon'}$, что при всех $n > n_{\epsilon'}$,

$$W(n; \varphi, l, r) \geq \bar{a} - 3\alpha - \epsilon', \quad (4)$$

где $\bar{a} = \max_i f(\mathbf{y}_{i,r}^{(i)})$ и максимум берется по всем наборам длины l . Вспомним, что функция $f(\mathbf{y}_l, r)$ есть сумма l функций, зависящих каждая от r элементов набора \mathbf{y}_l , и их можно представить как математические ожидания некоторого обобщенного в узком смысле ПНЗ. Согласно теореме 2 § 2 максимум $f(\mathbf{y}_l, r)$ достигается на наборе длины $l_0(k, r)$. Остается показать, как выбирается параметр r . Зададимся числом ϵ и положим r равным наименьшему из чисел r' , при которых функция $\alpha(r')$ из определения обобщенного ПНЗ меньше $\epsilon'/4$. Тогда (1) запишется

в виде

$$W(n; \varphi; l, r) > \max f(y_{l,r}) - \frac{3}{4}\varepsilon - \varepsilon'.$$

Полагая $\varepsilon' = \varepsilon/4$ при $n > n_{\varepsilon/4}$, имеем

$$W(n; \xi, l_0(k, r), r) \geq \max f(y_{l,r}) - \varepsilon.$$

Это завершает доказательство теоремы.

Полученный результат допускает усиление, если циклический выигрыш максимизировать не только по длинам и составам наборов управлений, но и по величинам параметра r .

Теорема 2. Для всякого обобщенного ПНЗ из класса $Q([0, A]; k)$ и любого $\varepsilon > 0$ существуют n_ε и параметр r^* циклического выигрыша такие, что автоматы $(CD)_{k,n}^{(l,r)}$ при всех $n > n_\varepsilon$ имеют предельный выигрыш:

$$W(n; \xi, l_0(k, r^*), r^*) > \lim_{r \rightarrow \infty} \max f(y_{l_0(k, r)}, r) - \varepsilon.$$

Доказательство. Выберем n_ε так же, как и в доказательстве теоремы 1, а $r = \max(r_1, r_2)$, где r_1 и r_2 выбраны из условий $\alpha(r_1) < \varepsilon/8$, а r_2 таково, что

$$|\max f(y_{l_0(k, r_2)}, r) - \lim_{r \rightarrow \infty} \max f(y_{l_0(k, r)}, r)| < \frac{\varepsilon}{8}.$$

Теперь утверждение теоремы непосредственно следует из теоремы 1.

Сделаем замечание относительно величины $\sup_{y_l, l, r} f(y_l, r)$,

приближение к которой выдвинуто целью управления обобщенными ПНЗ. Очевидно, справедливо неравенство

$$\sup_{y_l, l, r} f(y_l, r) \geq \lim_{r \rightarrow \infty} \max f(y_{l_0(k, r)}, r).$$

Согласно лемме из § 4 и теореме 1 (§ 4) автоматы $(CD)_{k,n}^{(l,r)}$ образуют ε -оптимальное семейство, ибо они обеспечивают достижение цели управления: получать в пределе средний выигрыш

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t > \sup_{(y^\infty)} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - \varepsilon$$

для каждого фиксированного ε и любого обобщенного ПНЗ класса $Q([0, A]; k)$.

Рассмотрим в заключение обобщенные в узком смысле ПНЗ и назначим целью управления — выполнение неравенства из предыдущего абзаца. Пусть фазовое пространство — отрезок $[0, A]$ числовой прямой. Условие затухания влияния предыстории процесса (выделяющее обобщенные ПНЗ) выполнено тривиальным образом. Поэтому семейство автоматов $(CD)_{k,n}^{l,r}$ служит ε -оптимальным семейством, причем наивыгоднейшая длина l_0 набора управлений указывается теоремой 2 (§ 2) для всех процессов класса $Q(X; k, r)$.

Легко убедиться, что это семейство ε -оптимально в сильном смысле

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \xi_t > \sup_{(y^\infty)} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - \varepsilon$$

в том же классе процессов.

ГЛАВА VIII

АЛГОРИТМЫ УПРАВЛЕНИЯ МАРКОВСКИМИ ПРОЦЕССАМИ

§ 1. Предварительные замечания

Рассматриваются объекты управления, принадлежащие классу управляемых марковских процессов со значениями в фазовом пространстве X (с σ -алгеброй \mathfrak{M}) и пространством управлений Y , которые описываются условными вероятностями (переходными функциями) $\mu(M|x, y)$, $M \in \mathfrak{M}$. На траекториях такого процесса ξ_t задан функционал $\varphi_t = \varphi(\xi_t, y^{t-1})$, у которого математическое ожидание $W_\sigma(t)$ существует и конечно при всех стратегиях σ из некоторого множества Σ . Цель управления относится к свойствам этого функционала. Сформулируем несколько типичных целей управления марковскими процессами. Начнем с экстремальных задач. Наиболее, быть может, распространенная цель состоит в требовании максимизировать

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E_\sigma \varphi(\xi_t, y_{t-1}).$$

В широких предположениях доказывается, что решая эту задачу стратегия стационарная марковская, т. е. однозначно определяется одной измеримой функцией $f: X \rightarrow Y$. Ее значение $f(x)$ в точке x указывает то действие y , которое должно быть совершено в момент t , если $\xi_t = x$. Простоты ради отождествим стратегию σ с функцией f , тогда можно определяемое стратегией действие в момент t записывать в виде $f(\xi_t)$ или $\sigma(\xi_t)$. Переходная функция тогда принимает форму $\mu(\cdot | x, \sigma)$.

Пусть Σ_{SM} — множество стационарных марковских стратегий. Допустим, что марковские процессы, порожденные управляемым марковским процессом ξ_t и любой стратегией $\sigma \in \Sigma_{SM}$, эргодические. Тогда существуют и не

зависят от начальных значений предельные распределения $\pi_\sigma(M)$.

Зададим на траекториях процесса функционал $\varphi_t = \varphi(\xi_t)$ — выигрыш в момент t — и введем предельный средний выигрыш в единицу времени

$$W(\sigma) = \int_{\mathbf{X}} \varphi(x) \pi_\sigma(dx).$$

К этой функции на множестве Σ относится следующая цель управления: максимизировать предельный средний выигрыш, т. е. найти стратегию σ_0 , удовлетворяющую условию

$$W(\sigma_0) = \max_{\sigma} W(\sigma). \quad (1)$$

В тех случаях, когда максимум $W(\sigma)$ не существует, приходится довольствоваться ϵ -оптимальностью, т. е. искать для каждого фиксированного $\epsilon > 0$ стационарную марковскую стратегию σ' , для которой

$$W(\sigma') > \sup_{\sigma} W(\sigma) - \epsilon. \quad (2)$$

Впрочем, иногда ϵ -оптимальность выдвигают в качестве цели управления и при существовании максимума среднего выигрыша $W(\sigma)$.

Рассмотрим теперь функционалы более сложного вида, которые зависят от значений марковского процесса более чем в один момент. В одном из вариантов функционалом служит $\varphi(\xi_{t+1}, \xi_t)$ — функция на паре состояний в соседние моменты времени (доход в результате перехода $\xi_t \rightarrow \xi_{t+1}$), а в другом функция $\varphi(\xi_{t+1}, \xi_t; \sigma)$, которая явным образом зависит от использованной стационарной марковской стратегии. Второй случай более общий и будем говорить о нем. Средний доход в состоянии x при стратегии σ определяется как математическое ожидание

$$g(x, \sigma) = \int_{\mathbf{X}} \varphi(z, x; \sigma) \mu(dz|x, \sigma),$$

и предельный средний доход равен

$$W(\sigma) = \int_{\mathbf{X}} g(x, \sigma) \pi_\sigma(dx).$$

В отношении таких функций цель заключается либо в максимизации предельного среднего дохода, либо в ϵ -оптимальности, — построении стратегии σ^* , для которой $W(\sigma^*) > \sup_{\sigma} W(\sigma) - \epsilon$. Обе постановки задачи сходны соответственно с целями (1) и (2).

Следуя терминологии гл. I (§ 4), перечисленные цели управления назовем *слабыми*. В некоторых случаях достижимость слабых целей влечет за собой достижимость сильных целей. Применительно к слабой цели (2) сильная ее модификация означает, что при всех $t > \tau_{\epsilon}$, где $P(\tau_{\epsilon} < \infty) = 1$, справедливо неравенство

$$\frac{1}{t} \sum_{l=1}^t \varphi(\xi_l) > \sup_{\sigma} W(\sigma) - \epsilon, \quad (3)$$

в котором ϵ фиксировано. Эта и подобные сильные цели вытекают из применимости усиленного закона больших чисел. Для конечных эргодических целей усиленный закон больших чисел справедлив, как это отмечалось в § 2 главы I.

Иной тип целей заключается в требовании, чтобы начиная с некоторого момента на траектории процесса выполнялось неравенство $\varphi_t > 0$ или, в несколько иной форме, $\overline{\lim}_{t \rightarrow \infty} \varphi_t > 0$. Другое представление подобной цели

относится к математическому ожиданию $W_{\sigma}(t) = E_{\sigma} \varphi_t$ по мере, порожденной стратегией σ , а именно, требуется, чтобы $W_{\sigma}(t) \leqslant 0$.

Ранее мы уже отмечали, что оптимизационные цели возможно задавать неравенствами. Так цель (3) записывается в виде

$$\frac{1}{t} \sum_{l=1}^t \varphi(\xi_l) + \epsilon - \sup_{\sigma} W(\sigma) > 0.$$

Небесполезно указать, что в типических «адаптивных» случаях такие функционалы не наблюдаются, потому что величина $\sup_{\sigma} W(\sigma)$ неизвестна.

σ

Конкретным примером цели управления, в которой выполнение (или невыполнение) неравенства наблюдаемо, служит «стабилизация» траектории марковского процесса. В случае нормированного фазового пространства X эта цель изображается в виде $\|\xi_t\| < R$, $R > 0$, при всех $t > \tau_R$ и заданном R . Другая цель управления требует устойчивости процесса в среднем, $\lim_{t \rightarrow \infty} E \|\xi_t\|^2 = 0$.

Стационарные марковские стратегии оказываются приводящими к цели не только для оптимизационных задач, но и для многих важных задач стабилизации и устойчивости (например, для линейных систем с постоянными коэффициентами).

В проблемах адаптивного управления марковскими процессами рассматриваются классы таких процессов с одними и теми же пространствами X и Y , но различными переходными функциями. Оказывается существенным, наблюдается ли сам управляемый марковский процесс или же некоторая функция от него. Причиной этого является то известное обстоятельство, что функция марковского процесса, вообще говоря, не есть марковский процесс.

Поэтому при адаптивной постановке задачи управления наблюдение не за самим процессом, а за функцией от него, часто делает почти бесполезной информацию о том, что в основе явления лежит марковский процесс. В описанной ситуации может оказаться более полезным решать задачу управления классом процессов, который получен из данного класса марковских процессов соответствующим преобразованием.

Ниже всегда предполагается, что значения управляемого марковского процесса наблюдаются. Интересующий нас функционал, к которому относится цель управления, может либо наблюдаваться, либо, если задан вид его зависимости от процесса, вычисляться системой управления.

Почти при всех целях адаптивного управления марковскими процессами приходится ограничиваться процессами без несущественных состояний и всего с одним эргодическим классом. В противном случае цель обычно не достигается.

§ 2. ε -оптимальные семейства для эргодических марковских процессов

Символ ξ_t относится здесь к однородным эргодическим марковским процессам с фазовым пространством $X = [0, 1]$ и конечным пространством управлений $Y = \{y_1, \dots, y_k\}$. Ограничиваая стратегии множеством $\Sigma_M = \{\sigma\}$ стационарных марковских, определим соответствующие предельные средние выигрыши $W(\sigma)$ и сформулируем цель управления как ε -оптимальность в слабом смысле: построить семейство автоматов A_n , которые при всех достаточно больших t обеспечивают выполнение неравенства

$$E\xi_t > \sup_{\sigma \in \Sigma} W(\sigma) - \varepsilon,$$

в котором математическое ожидание слева берется по стратегии, реализуемой соответствующим автоматом этого семейства.

Предположенная эргодичность процесса приводит к ε -оптимальности в сильном смысле:

$$\frac{1}{t} \sum_{i=1}^t \xi_i > \sup_{\sigma \in \Sigma} W(\sigma) - \varepsilon$$

при всех $t > \tau_\varepsilon$.

Наложим дополнительное ограничение на переходные функции рассматриваемых здесь классов:

существуют плотности $p(z | x, \sigma)$ переходных функций $\mu(\cdot | x, \sigma)$, удовлетворяющие при всех z, x, σ неравенствам

$$0 < c \leq p(z | x, \sigma) \leq C < \infty.$$

Определенный таким образом класс марковских процессов обозначим $M(c, C; k)$.

Установим некоторые вспомогательные факты.

Пусть $\eta_i^{(1)}, \eta_i^{(2)}$ — марковские процессы со значениями на отрезке $[0, 1]$ и плотностями $p_i(z | x)$, $i=1, 2$, переходных функций таких, что

$$0 < c \leq p_i(z | x) \leq C < \infty. \quad (1)$$

Можно показать, что для каждого из этих процессов существуют предельные распределения π_1 и π_2 , имеющие

плотности π_1^0 и π_2^0 . Случайные величины с этими распределениями обозначим ζ_1 и ζ_2 .

Л е м м а 1. *При высказанных предположениях на процессы $\eta_t^{(1)}$ и $\eta_t^{(2)}$ и выполнении равенства $p_1(z|x) = p_2(z|x)$, справедливого при всех z , кроме измеримого множества Γ меры $\text{mes } \Gamma$, имеет место оценка*

$$|\mathbf{E}\zeta_1 - \mathbf{E}\zeta_2| \leq \frac{C^2}{c} \text{mes } \Gamma.$$

Д о к а з а т е л ь с т в о. Для плотностей π_i^0 , $i=1, 2$, выполнены очевидные равенства

$$\pi_i^0(x) = \int_0^1 p_i(x|z) \pi_i^0(z) dz. \quad (2)$$

В пространстве L_1 интегрируемых на $[0, 1]$ функций зададим два ограниченных оператора

$$T_i f(x) = \int_0^1 \bar{p}_i(x|z) f(z) dz, \quad i=1, 2, \quad (3)$$

где принято сокращение $\bar{p}_i = p_i - c$.

Покажем, что эти операторы сжимающие. В самом деле,

$$\begin{aligned} \|T_i f\| &= \int_0^1 \left| \int_0^1 \bar{p}_i(x|z) f(z) dz \right| dx \leq \\ &\leq \int_0^1 \int_0^1 |\bar{p}_i(x|z)| |f(z)| dz dx = (1-c) \|f\|. \end{aligned}$$

Значит, каждый из них имеет единственную собственную функцию. Для T_i , согласно (2) и (3), ей служит соответствующая плотность предельного распределения π_i^0 .

Теперь оценим $\|\pi_1^0 - \pi_2^0\|$:

$$\|\pi_1^0 - \pi_2^0\| = \|T_1 \pi_1^0 - T_2 \pi_2^0\| \leq \|T_1 \pi_1^0 - T_2 \pi_1^0\| + \|T_2 \pi_1^0 - T_2 \pi_2^0\|.$$

Для первого слагаемого в правой части имеем

$$\|T_1\pi_1^0 - T_2\pi_1^0\| =$$

$$\begin{aligned} &= \int_0^1 \left| \int_0^1 p_1(x|z) \pi_1^0(z) dz - \int_0^1 p_2(x|z) \pi_1^0(z) dz \right| dx = \\ &= \int_0^1 \int_0^1 |p_1(x|z) - p_2(x|z)| \pi_1^0(z) dz dx \leq C^2 \operatorname{mes} \Gamma, \end{aligned}$$

а для второго слагаемого учитываем, что T_2 — сжимающий оператор

$$\|T_2(\pi_1^0 - \pi_2^0)\| \leq (1 - c) \|\pi_1^0 - \pi_2^0\|_{L_1}.$$

Отсюда получаем

$$\|\pi_1^0 - \pi_2^0\| \leq C^2 \operatorname{mes} \Gamma + (1 - c) \|\pi_1^0 - \pi_2^0\|.$$

Это приводит к утверждению леммы.

Возвращаясь к управляемым марковским процессам, заметим прежде всего, что каждая стратегия из Σ определяет разбиение $\Delta = (\Delta_1, \dots, \Delta_k)$ отрезка $[0, 1]$ на измеримые подмножества $\Delta_i = \{x : \sigma(x) = y_i\}$, $i = 1, \dots, k$. Нам потребуется оценивать «расстояния» между разбиениями отрезка $[0, 1]$. Введем псевдометрику. Пусть $\Delta' = (\Delta'_1, \dots, \Delta'_n)$, $\Delta'' = (\Delta''_1, \dots, \Delta''_m)$ — два разбиения, тогда

$$\rho(\Delta', \Delta'') = 1 - \sum_{i=1}^m \max_j \operatorname{mes}(\Delta'_j \cap \Delta''_j).$$

Легко проверить, что не выполнено свойство симметрии $\rho(\Delta', \Delta'') \neq \rho(\Delta'', \Delta')$. Далее, если разбиение Δ''' получено из Δ'' объединением некоторых подмножеств, то при любом Δ'

$$\rho(\Delta', \Delta'') \leq \rho(\Delta', \Delta''').$$

Л е м м а 2. *Каково бы ни было измеримое разбиение Δ' отрезка $[0, 1]$ на k подмножеств и число $\epsilon > 0$, найдется разбиение Δ'' этого отрезка на равные полуинтервалы такое, что $\rho(\Delta', \Delta'') \leq \epsilon$.*

Доказательство. Пусть $\Delta' = (\Delta'_1, \dots, \Delta'_k)$ — заданное измеримое разбиение и ϵ — фиксированное число. Для

каждого множества Δ'_i существует такое целое m_i , что из разбиения $[0, 1]$ на 2^{m_i} полуинтервалов $[0, 2^{-m_i}), [2^{-m_i}, 2 \cdot 2^{-m_i}), \dots, [n2^{-m_i}, (n+1)2^{-m_i}, \dots, [(2^{m_i}-1)2^{-m_i}, 1]$ можно выделить систему полуинтервалов $J(m_i)$, обладающую свойством

$$\text{mes}(\Delta'_i s J(m_i)) < \frac{\varepsilon}{k^2}, \quad (4)$$

где s — символ симметрической разности множеств. Положим $m = \max_i m_i$ и $\Delta''(m)$ — разбиение отрезка на 2^m полуинтервалов. Для всякого i из $\Delta''(m)$ можно выделить систему полуинтервалов J_i , для которой справедливо (4). Это означает

$$|\text{mes } \Delta'_i - \text{mes } J_i| < \frac{\varepsilon}{k^2}.$$

Из того, что элементы разбиения Δ' не пересекаются, вытекает

$$\text{mes}(J_i \cap J_j) < \frac{\varepsilon}{k^2}. \quad (5)$$

Введем еще одно разбиение отрезка $[0, 1]$ на $k+1$ множества, которые определяются равенствами

$$\Delta''_i = J_i \setminus \bigcup_{j \neq i} J_j, \quad i = 1, \dots, k; \quad \Delta''_{k+1} = [0, 1] \setminus \bigcup_{i=1}^k \Delta''_i,$$

впрочем, среди них могут оказаться пустые. Из определения этого разбиения, обозначим его Δ''' , и неравенств (4), (5) для любого $i \leq k$ следует

$$\text{mes } \Delta'_i - \frac{\varepsilon}{k} < \text{mes}(\Delta'_i \cap \Delta''_i) \leq \text{mes } \Delta'_i + \frac{\varepsilon}{k}, \quad i = 1, \dots, k. \quad (6)$$

Очевидно, кроме того, что $\text{mes } \Delta'''_{k+1} < \varepsilon/k$. Отсюда и из (6) находим

$$\begin{aligned} \rho(\Delta', \Delta'') &= 1 - \sum_{i=1}^{k+1} \max_j \text{mes}(\Delta'_j \cap \Delta''_i) \leq \\ &\leq 1 - \sum_{i=1}^k \left(\text{mes } \Delta'_i - \frac{\varepsilon}{k} \right) = \varepsilon. \end{aligned}$$

Разбиение Δ''' получено объединением некоторых подмножеств из $\Delta''(m)$, поэтому высказанное перед леммой 2 свойство расстояния ρ приводит к требуемому неравенству $\rho(\Delta', \Delta''(m)) \leq \rho(\Delta', \Delta''') < \varepsilon$.

Полученный результат можно перефразировать в терминах стратегий.

Сначала введем еще одно понятие: стратегия называется *кусочно постоянной*, если порожденное ею разбиение отрезка $[0, 1]$ состоит из конечного числа полуинтервалов.

Л е м м а 2'. Для любой марковской (измеримой) стратегии $\sigma(x)$ и любого $\varepsilon > 0$ существует кусочно постоянная стратегия $\sigma'(x)$ такая, что мера множества $\{x : \sigma(x) \neq \sigma'(x)\}$ не превосходит ε .

Высказанные утверждения позволяют прийти к следующей теореме.

Т е о р е м а 1. Для всякого процесса из класса $M(c, C; k)$, произвольной марковской (измеримой) стратегии $\sigma(x)$ и любого $\varepsilon > 0$ существует кусочно постоянная стратегия $\sigma'(x)$ такая, что для случайных величин ζ' и ζ'' с распределениями π_{σ} и $\pi_{\sigma'}$, соответственно, имеет место неравенство

$$|E\zeta' - E\zeta''| < \varepsilon.$$

Д о к а з а т е л ь с т в о. Зададимся числом $\varepsilon > 0$ и выберем в классе $M(c, C; k)$ процесс ξ_t . Согласно лемме 2' для стратегии $\sigma(x)$ подберем такую кусочно постоянную стратегию $\sigma'(x)$, что $\text{mes}\{\{x : \sigma(x) \neq \sigma'(x)\} < \frac{c}{C^2} \varepsilon$. Значит, отвечающие этим стратегиям плотности переходной функции различаются на множестве Γ , мера которого менее $\frac{c}{C^2} \varepsilon$. Остается воспользоваться леммой 1, из которой непосредственно следует доказываемое неравенство.

Введем множество Σ_m марковских кусочно постоянных стратегий, порождаемых разбиением $\Delta = (\Delta_1, \dots, \Delta_{2^m})$ отрезка $[0, 1]$ на равные интервалы длины 2^{-m} . Всего в множестве Σ_m содержится k^{2^m} элементов.

Обратимся теперь к построению требуемого ε -оптимального семейства автоматов. Изберем для них обозначение $(MD)_{k, n}^{(l, m)}$. Входными сигналами служат значения управ-

ляемого марковского процесса ξ_t , т. е. числа из $[0, 1]$. Выходные сигналы — элементы $Y = \{y_1, \dots, y_k\}$. Наконец, множество состояний представляет собой множества пар $(\sigma^{(i)}, r)$, где $\sigma^{(i)} \in \Sigma_m$, а $r=1, \dots, n$ — номер состояния « i -й ветви». При каждом i состояния $(\sigma^{(i)}, 1), (\sigma^{(i)}, 2), \dots, (\sigma^{(i)}, n)$ связаны переходами так же, как и в автомате $D_{k,n}$. Отличие заключается в том, что в каждом состоянии автомат совершают действия, определяемые соответствующей стратегией, l раз и вычисляет усредненный выигрыш

$$h_t(\sigma^{(i)}, l) = \frac{1}{l} \sum_{\mu=1}^l \xi_{t+\mu}.$$

Эта величина принадлежит, разумеется, отрезку $[0, 1]$. С вероятностью $h_t(\sigma^{(i)}, l)$ это значение принимается «поощрением», т. е. приемлемым, и с дополнительной вероятностью — «наказанием». В случае поощрения состояние $(\sigma^{(i)}, r)$ замещается на $(\sigma^{(i)}, n)$, а в случае наказания на $(\sigma^{(i)}, r-1)$. Из $(\sigma^{(i)}, 1)$ при наказании совершается переход в $(\sigma^{(j)}, 1)$, где индекс j новой кусочно постоянной стратегии выбирают, например, равновероятно.

Согласно усиленному закону больших чисел

$$\lim_{l \rightarrow \infty} h_t(\sigma, l) = E\zeta_\sigma$$

и, кроме того, в силу ограниченности процесса,

$$\lim_{l \rightarrow \infty} E h_t(\sigma, l) = E\zeta_\sigma$$

для всякой марковской стратегии σ из класса Σ_m .

Определим предельный средний выигрыш автомата $(MD)_{k,n}^{(l,m)}$ равенством

$$W(n; l, m) = \lim_{t \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l E\xi_{t+i} = \lim_{t \rightarrow \infty} E\xi_t,$$

где математическое ожидание берется по мере, индуцированной реализуемой автоматом стратегией. Легко видеть, что оба предела существуют.

Будем рассматривать $(MD)_{k,n}^{(l,m)}$ только в моменты времени, кратные l , когда подсчитывается усредненный выигрыш. Тогда этот автомат можно интерпретировать как автомат $\tilde{D}_{v,n}$ ($v = k^{2m}$), управляющий процессом $h_t(l) = \frac{1}{l} \sum_{\mu=1}^l \xi_{t+\mu}$. Выходными сигналами его служат стратегии из Σ_m . Предельный средний выигрыш этого автомата выражается так:

$$W(\tilde{D}_{v,n}) = \lim_{t \rightarrow \infty} h_t(l).$$

Л е м м а 3. *Процесс $h_t(l)$ является П-процессом.*

Д о к а з а т е л ь с т в о. Распределения вероятностей этого процесса определяются, вообще говоря, всей его предысторией. Из сделанных предположений о плотностях переходных функций класса $M(c, C; k)$, влекущих за собой эргодичность рассматриваемых цепей, можно утверждать, что для любых t_1, t_2 и всякой стратегии σ из Σ_m

$$|\mathbf{E}h_{t_1}(\sigma, l) - \mathbf{E}h_{t_2}(\sigma, l)| < \beta(l), \quad (7)$$

где $\beta(l)$ экспоненциально быстро стремится к 0 (когда $l \rightarrow \infty$) равномерно по всем предысториям процесса ξ_t (и, следовательно, h_t). Лемма доказана.

В следующей теореме указано интересующее нас свойство семейства автоматов $(MD)_{k,n}^{(l,m)}$ по отношению к классу управляемых марковских процессов M_k , определяемому равенством

$$M_k = \bigcup_{\substack{c, C \\ (0 < c < C < \infty)}} M(c, C; k).$$

Т е о р е м а 2. *Автоматы $(MD)_{k,n}^{(l,m)}$ образуют ε — оптимальное семейство для класса M_k управляемых марковских процессов, т. е. для каждого $\xi_t \in M_k$ и любого $\epsilon > 0$ найдутся такие целые числа $n_\epsilon, l_\epsilon, m_\epsilon$, что при всех $n > n_\epsilon, l > l_\epsilon, m > m_\epsilon$ выполнены неравенства*

$$W(n; l, m) > \sup_{\sigma} \mathbf{E}\xi_{\sigma} - \epsilon.$$

Доказательство. Зафиксируем $\varepsilon > 0$; найдется такая марковская (измеримая) стратегия σ^* , что

$$E\xi_{\sigma^*} > \sup_{\sigma} E\xi_{\sigma} - \frac{\varepsilon}{\sigma}.$$

Согласно теореме 1 существует такое целое число m_{ε} , что в Σ_m , $m > m_{\varepsilon}$, существует стратегия $\sigma^{(m)}$, для которой

$$E\xi_{\sigma^{(m)}} > E\xi_{\sigma^*} - \frac{\varepsilon}{\sigma}.$$

Основываясь на лемме 3, выберем l_{ε} из условия $\beta(l_{\varepsilon}) < \varepsilon/6$. Тогда при всех $l > l_{\varepsilon}$ имеем

$$|Eh_{t_1}(\sigma, l) - Eh_{t_r}(\sigma, l)| < \frac{\varepsilon}{\sigma}.$$

Из определений величины $W(n; l, m)$ и $W(\tilde{D}_n)$ леммы 3 и леммы из § 5 гл. VII следует, что существует целое число n_{ε} , обладающее свойством: при всех $n > n_{\varepsilon}$ справедливы неравенства

$$W(n; l, m) = W(\tilde{D}_n) > \max_{\sigma \in \Sigma_m} E\xi_{\sigma} - 3\beta(l) - \frac{\varepsilon}{\sigma}.$$

Таким образом, для любого $\varepsilon > 0$ указана требуемая теоремой тройка чисел $n_{\varepsilon}, l_{\varepsilon}, m_{\varepsilon}$. Этим завершено доказательство теоремы.

Полученный результат может быть записан в виде равенства

$$\lim_{n, l, m \rightarrow \infty} W(n; l, m) = \sup_{\sigma} E\xi_{\sigma},$$

справедливого для каждого управляемого марковского процесса класса M_k . Нетрудно, далее, убедиться, что изученные здесь автоматы образуют ε -оптимальное в сильном смысле семейство.

Таким образом, решена поставленная в начале параграфа задача. Следует обратить внимание на сложность изученных автоматов, на весьма большое число содержащихся в них состояний, если автоматы должны обеспечить малое значение погрешности ε . Одна из причин тому — бесконечность фазового пространства процесса, а значит, и бесконечность множества стратегий. Струк-

тура адаптивных систем для марковских процессов с конечным числом состояний оказывается достаточно простой. Ниже излагается один из вариантов построения ε -оптимальных семейств конечных автоматов для конечных марковских цепей.

Пусть $\{X, \mathcal{P}^{(y)}, Y\}$ — конечная управляемая марковская цепь, где $X = \{x_1, \dots, x_m\}$, $x_i \in [0, 1]$ — состояния, $\mathcal{P}^{(y)} = \|p_{ij}(y)\|$ — матрицы вероятностей переходов из i -го состояния в j -е под влиянием управления $y \in Y = \{y_1, \dots, y_k\}$. Допустим, при $i, j \in [1, \dots, m]$, $y \in Y$

$$p_{ij}(y) > 0.$$

Это значит, что все марковские цепи, отвечающие стационарным марковским стратегиям, эргодические. Символ y мы вводим для стационарной марковской стратегии

$$y = (y(x_1), y(x_2), \dots, y(x_m)),$$

которой отвечает однородная цепь с матрицей вероятностей переходов

$$\mathcal{P}^{(y)} = \|p_{ij}(y(x_i))\|.$$

Свойство эргодичности рассматриваемых цепей означает, что при всякой стратегии y с вероятностью 1 существует и не зависит от начального состояния предел

$$\lim_{l \rightarrow \infty} \frac{1}{l} \sum_{t=1}^l \xi_t = E\xi_y.$$

Величину $h_t(l) = \frac{1}{l} \sum_{\mu=1}^l \xi_{t+\mu}$ назовем *усредненным выигрышем*.

Выскажем цель управления — построить ε -оптимальное в слабом смысле семейство конечных автоматов, т. е. чтобы при каждом фиксированном $\varepsilon > 0$ и достаточно больших t выполнялось неравенство

$$Eh_t(l) > \max_y E\xi_y - \varepsilon.$$

Ясно, что обеспечивающая эту цель система является ε -оптимальной в сильном смысле: при всех $t > \tau_\varepsilon$

$$\frac{1}{t} \sum_{i=1}^t \xi_i > \max_y E \zeta_y - \varepsilon.$$

Средством достижения этой цели мы изберем автоматы $(Md)_{k,n}^{(l)}$, определяемые сходным образом с автоматами $(MD)_{k,n}^{(l,m)}$. Их входными сигналами служат значения управляемой цепи, т. е. элементы множества X . Состояния разбиваются на k^n ветвей $((y, r), r=1, 2, \dots, n)$, где n — глубина памяти автомата. В каждом состоянии l тактов подряд совершаются действия, определяемые стратегией y , и подсчитывается усредненный выигрыш $h_t(l)$. Ему с вероятностью $h_t(l)$ сопоставляем «поощрение» и с дополнительной вероятностью — «наказание». При поощрении состояние (y, r) замещается на (y, n) , а при наказании — на $(y, r-1)$. Из состояния $(y, 1)$ с одинаковыми вероятностями совершается переход в одно из первых состояний, отвечающих любой из возможных k^n стратегий.

Итак, автоматы $(Md)_{k,n}^{(l)}$ реализуют стратегию управления марковскими цепями, которая порождает некоторую меру. Вычисленное по этой мере математическое ожидание $E \xi_t$ имеет предел по времени. В этом несложно убедиться, опираясь на очевидную эргодичность ассоциированной марковской цепи (см. гл. I), у которой множеством состояний является $X \times S_{n,l}$, где $S_{n,l}$ — множество состояний автомата. Обозначим предельный средний выигрыш автомата $W(n, l)$. Обозначим через $M(X, k)$ класс введенных здесь конечных эргодических марковских цепей с одним и тем же фазовым пространством.

Теорема 3. Автоматы $(Md)_{k,n}^{(l)}$ образуют по индексам l и n ε -оптимальное семейство.

Доказательство протекает аналогично доказательствам предыдущих теорем. Надо лишь убедиться, что процесс $h_t(l)$ является Π -процессом, и снова воспользоваться леммой из § 5 гл. VII.

§ 3. Алгоритмы управления конечными марковскими цепями с доходами

Пусть $S = \{s_1, \dots, s_m\}$ — множество состояний, $Y = \{y_1, \dots, y_k\}$ — множество управлений. Определены семейства матриц $\mathcal{P}^{(y)} = \|p_{ij}(y)\|$ вероятностей переходов и $R^{(y)} = \|r_{ij}(y)\|$ доходов (за переход из i -го состояния в j -е при управлении y).

Рассматривается класс $MD(S, Y)$ эргодических марковских цепей с доходами $(S, Y, \mathcal{P}^{(y)}, R^{(y)})$ с одними и теми же множествами S и Y . Символ y сохраним за стационарными марковскими стратегиями ($y = y(s)$).

Сформулируем цели управления такими цепями. Одношаговой прибылью в i -м состоянии за действие y называется величина

$$q(i, y) = \sum_{j=1}^m p_{ij}(y) r_{ij}(y),$$

равная математическому ожиданию дохода за переход из i -го состояния под действием y . Через $\pi(y) = (\pi_1(y), \dots, \dots, \pi_m(y))$ обозначим предельное распределение состояний однородной эргодической марковской цепи $(S, \mathcal{P}^{(y)})$. Предельным средним (одношаговым) доходом назовем

$$W(y) = \sum_{i=1}^m q(i, y(s_i)) \pi_i(y),$$

где $y(s_i)$ — управление, которое стратегия y сопоставила состоянию s_i , $i = 1, \dots, m$.

Одна из выдвигаемых целей управления заключается в ϵ -оптимальности в сильном смысле, т. е. в обеспечении для каждого процесса из класса $MD(S, Y)$ выполнения неравенств (φ_l — доход в момент l)

$$\frac{1}{t} \sum_{l=1}^t \varphi_l > \max_y W(y) - \epsilon$$

для любого фиксированного $\epsilon > 0$ при всех $t > \tau_\epsilon$, причем $P(\tau_\epsilon < \infty) = 1$. Такую цель должно гарантировать ϵ -оптимальное семейство адаптивных систем.

Другая цель управления состоит в асимптотической оптимальности в сильном смысле: для каждого процесса из класса должны выполняться неравенства

$$\frac{1}{t} \sum_{l=1}^t \varphi_l > \max_y W(y) - \epsilon$$

для произвольного $\epsilon > 0$ при всех $t > \tau_\epsilon$ ($P(\tau_\epsilon < \infty) = 1$). Эту цель обеспечивает асимптотически оптимальная адаптивная система.

Описанию конструкций соответствующих адаптивных систем предпоследним замечания о способах достижения целей в случае, если марковская цепь с доходами задана полностью, т. е. известны все $2k$ матриц $\mathcal{P}^{(y)}$ и $R^{(y)}$. Простейшим в идейном плане, хотя и громоздким практически, является следующий метод перебора. Для каждой из k^m стратегии вычисляется предельный средний выигрыш $W(y)$, а затем выбирается наибольший выигрыш \bar{W} . Отвечающая ему стратегия y_0 ($W(y_0) = \bar{W}$) — искомая оптимальная стратегия, впрочем, их может быть несколько. Другой способ — итеративный — менее трудоемок и будет описан ниже.

Излагаемые ниже алгоритмы управления цепями с доходами основаны на перечисленных способах. Начнем с тех, которые опираются на метод перебора. Мы преследуем обе сформулированные цели управления. Обеспечивающие их алгоритмы сходные и поэтому они описываются одновременно.

В основе первого алгоритма лежат те же принципы, что и для-автоматов. Не стремясь к общности, которая может затмнить идеи, воспользуемся схемой автоматов типа G (§ 6 гл. III). Более того, сохраним использованные там обозначения.

Перенумеруем как-нибудь, например лексикографически, все k^m марковских правил совершения действий y_1, \dots, y_{k^m} и сопоставим каждому из них пару функций $(N_i(t), V_i(t))$. Иногда удобно отдельно записывать функции N_i и V_i , которые образуют k^m -мерные векторы

$$\mathbf{N}(t) = (N_1(t), \dots, N_{k^m}(t)), \quad \mathbf{V}(t) = (V_1(t), \dots, V_{k^m}(t)).$$

Заданию этих функций предпоследнее определение *i*-подцепи, т. е. подпоследовательности

$$\xi_{t_1}, \xi_{t_1+1}, \xi_{t_2}, \xi_{t_2+1}, \dots, \xi_{t_y}, \xi_{t_y+1}, \xi_{t_y+1+1}, \dots, \quad (1)$$

выделенной из последовательности пар (ξ_t, ξ_{t+1}) состояний марковской цепи двумя условиями

$$\xi_{t_y+1} = \xi_{t_y+1}, \quad y_{t_y} = y^i(\xi_{t_y}). \quad (2)$$

Они означают, что первое состояние ξ_{t_y+1} очередной пары $(\xi_{t_y+1}, \xi_{t_y+1+1})$ совпадает со вторым состоянием предыдущей пары (ξ_{t_y}, ξ_{t_y+1}) , которое появилось в результате перехода $\xi_{t_y} \rightarrow \xi_{t_y+1}$ под действием управления, сопоставленного *i*-стратегии исходному состоянию ξ_{t_y} .

Подпоследовательность (1) пар (ξ_{t_y}, ξ_{t_y+1}) представляет собой эргодическую марковскую цепь с предельными вероятностями состояний $(\pi_j(y_i) p_{j,l}(y^{(i)}(s_j)), j, l = 1, \dots, m)$. Доходы $r_{j,l}(y^{(i)}(s_j))$ являются на ней функционалами.

Значение функции $N_i(t)$ в момент t равно числу переходов вида (1) за время t , т. е. количеству применений правила y_i . Функция $V_i(t)$ — эмпирический средний выигрыш *i*-го правила — служит оценкой $W(y_i)$ и выражается формулой

$$V_i(t) = \frac{1}{N_i(t)} \sum_{y=1}^{N_i(t)} \varphi_{t,y}, \quad (3)$$

где $\varphi_{t,y} = r_{\xi_{t_y}, \xi_{t_y+1}}(y^{(i)}(\xi_{t_y}))$ — случайная величина — доход, полученный за переход $\xi_{t_y} \rightarrow \xi_{t_y+1}$ при использовании в состоянии ξ_{t_y} того управления, которое приписывает этому состоянию правило y_i .

Из сказанного ясны правила трансформации функций $N_i(t)$ и $V_i(t)$: если в момент t применялось действие *i*-го правила, то к $N_i(t-1)$ прибавляется 1, а $V_i(t)$ пересчитывается очевидным из формулы (3) образом. Необходимо заметить, что в каждый момент времени преобразуется не одна пара функций (N_i, V_i) , а k^{m-1} пар, соответствующих всем тем из k^{m-1} правил, у которых состоянию ξ_t , отвечает одно и то же управление $y(\xi_t)$.

Эргодичность всех цепей $(S, \mathcal{P}^{(y)})$ и усиленный закон больших чисел приводят к тому, что если $N_i(t)$ неогра-

ниченно растут, $V_{i_j}(t)$ оказываются состоятельными и несмещеными оценками величин $W(y_i)$, т. е.

$$\mathbb{P}(\lim_{t \rightarrow \infty} V_{i_j}(t) = W(y_i), i = 1, \dots, k^m) = 1. \quad (4)$$

Через $B(t)$ обозначим стохастическую $m \times k$ -матрицу $B(t) = \|B_{ij}(t)\|$ с элементами $B_{ij}(t)$, означающими вероятности совершить действие y_j в момент t , если цепь находится в состоянии s_i . Предполагается $B_{ij}(t) > 0$ при всех i, j, t . Начальным значением матрицы $B(0)$ служат элементы $B_{ij}(0) = 1/k$. Ниже сформулируем правило ее преобразования. Необходимо задать числовую последовательность $\delta(n)$, подчиненную следующим условиям:

$$0 < \delta(n) < 1/2, \quad \lim_{n \rightarrow \infty} \delta(n) = \delta,$$

где δ либо положительно, либо равно нулю, причем $\delta(n)$ монотонно убывают. Выполняемая процедура такова: из эмпирических средних выигрышей строится вариационный ряд

$$V_{i_1}(t) \geqslant V_{i_2}(t) \geqslant \dots \geqslant V_{i_{k^m}}(t).$$

Соответственно упорядочим правила $y_{i_1}, y_{i_2}, \dots, y_{i_{k^m}}$. В n -м преобразовании отбираются «лучшие» правила (с наибольшими эмпирическими средними выигрышами) в количестве $x_n = \max(1, [\theta^n \cdot k^m])$. Через G_i обозначим группу действий, сопоставляемых состоянию s_i цепи «лучшими» правилами, а через $|G_i|$ — количество элементов в ней. До тех пор, пока все числа $|G_i|$ равны k , матрица $B(t)$ остается неизменной. Впервые строка матрицы, соответствующая i -му состоянию, может смениться, когда окажется $|G_i| < k$. Если $|G_i| < k$, то в моменты преобразования матрицы $B(t)$ меняется одна ее строка, номер которой равен номеру текущего состояния цепи. Элементы этой строки в момент n -го преобразования становятся равными

$$B_{ij}(t_n) = \begin{cases} \frac{1 - \delta(n)}{|G_i|}, & j \in G_i, \\ \frac{\delta(n)}{k - |G_i|}, & j \notin G_i, \end{cases} \quad |G_i| < k.$$

После того как окажется $|G_i|=1$, строки матрицы $B(t)$ преобразуются по формуле

$$B_{ij}(t_n) = \begin{cases} 1 - \delta(n), & j \in G_i, \\ \frac{\delta(n)}{k-1}, & j \notin G_i, \end{cases} \quad |G_i| = 1.$$

Назовем описанный алгоритм *GM-алгоритмом*.

Перейдем к исследованию *GM-алгоритмов* в предположении, что матрица $B(t)$ преобразуется в моменты, когда возрастает на единицу величина

$$n^0(t) = \min_i N_i(t)$$

(после совершения наименее часто использовавшегося до момента t правила), и в монотонно убывающей последовательности $\delta(n)$ содержится конечное количество различных чисел, последним из которых является число $\delta > 0$. Будем их обозначать $GM(n^0, \delta)$ -алгоритмами. Очевидно, они представляют собой стратегии для класса $MD(S, Y)$ цепей с доходами.

Теорема 1. Семейство $GM(n^0, \delta)$ -алгоритмов является ε -оптимальным в сильном смысле семейством адаптивных систем в классе $MD(S, Y)$, т. е. для любого процесса из класса и каждого ε найдется такое δ_ε , что при любом $\delta < \delta_\varepsilon$ существует $\tau_\varepsilon(\delta)$, при котором для всех $t > \tau_\varepsilon(\delta)$ имеют место неравенства

$$\frac{1}{t} \sum_{i=1}^t \varphi_i > \max_y W(y) - \varepsilon.$$

Доказательство. Из того, что в каждом состоянии марковской цепи вероятности выбора всех действий не меньше положительного числа $\delta/(k-1)$, заключаем

$$\mathbb{P}(\lim_{t \rightarrow \infty} N_i(t) = \infty, i = 1, \dots, k^m) = 1.$$

Отсюда сразу вытекает состоятельность оценок $V_i(t)$ и справедливость равенства (4). Следовательно, наступит такой момент, начиная с которого эмпирические средние

выигрыши $V_i(t)$ окажутся упорядоченными в вариационном ряде так же, как и средние выигрыши $W(y_i)$. Допустим (без ограничения общности), что оптимальное правило одно и им является y_1 . Спустя случайное время τ , причем $P(\tau = \infty) = 0$, все группы G_i будут содержать ровно по одному элементу. Этим элементом служит действие, которое занимает i -компоненту в векторе y_1 . Вероятность выбора такого действия становится*) равной $1 - \delta$.

Остается показать, что рассматриваемые алгоритмы порождают ε -оптимальное семейство. Выпишем суммарный выигрыш за время $t > \tau$

$$V(t) = \frac{1}{t} \sum_{j=1}^t \varphi_j = \frac{1}{t} \sum_{j=1}^{\tau} \varphi_j + \frac{t-\tau}{t} \frac{1}{t-\tau} \sum_{j=\tau+1}^t \varphi_j.$$

При $t \rightarrow \infty$ первое слагаемое справа стремится к 0 и $\frac{t-\tau}{t} \rightarrow 1$. Изучим величину

$$\frac{1}{t-\tau} \sum_{j=\tau+1}^t \varphi_j = \sum_{j=1}^m \frac{l_j(t)}{t-\tau} U_j(t)$$

для $t > \tau$, где $l_j(t)$ — число попаданий цепи в состояние s_j за время $(\tau, t]$, а величина

$$U_j(t) = \frac{1}{l_j(t)} \sum_{i=1}^{l_j(t)} \varphi_i$$

равна эмпирическому среднему выигрышу в состоянии s_j за время $(\tau, t]$.

После достижения δ -оптимального режима (т. е. при $t > \tau$) переходы между состояниями однородной марковской цепи определяются матрицей вероятностей перехода $\mathcal{P}(\delta) = \|p_{ij}^\delta\|$, элементы которой выражаются формулой

$$p_{ij}^\delta = (1 - \delta) p_{ij}(y^{(1)}(s_i)) + \frac{\delta}{k-1} \sum_{\substack{n=1 \\ y_n \neq y^{(1)}(s_i)}}^k p_{ij}(y_n).$$

*) Ранее (в § 5 гл. III) это свойство было названо δ -оптимальностью. Поэтому реализующая $GM(n^0, \delta)$ -алгоритм адаптивная система может именоваться δ -оптимальной.

При достаточно малых δ цепи $(S, \mathcal{P}(\delta))$ эргодические. Предельные вероятности состояний π_i^δ удовлетворяют алгебраической системе

$$\sum_{i=1}^m p_{ij}^\delta \pi_i^\delta = \pi_j^\delta, \quad j = 1, \dots, m,$$

$$\pi_1^\delta + \dots + \pi_m^\delta = 1.$$

Отсюда легко получить

$$\pi_i^\delta = \pi_i(y_i) + c_i \delta, \quad i = 1, \dots, m,$$

где $\sum_{i=1}^m c_i = 0$. Значит, с вероятностью 1 справедливы равенства

$$\lim_{t \rightarrow \infty} \frac{l_j(t)}{t - \tau} = \pi_j(y_1) + c_j \delta, \quad j = 1, \dots, m. \quad (5)$$

К каждой последовательности $U_j(t)$ применим усиленный закон больших чисел, справедливый для конечных однородных эргодических цепей. Следовательно, при достаточно малых δ имеем с вероятностью единица

$$\lim_{t \rightarrow \infty} U_j(t) = E_\delta \varphi_{t,y}, \quad j = 1, \dots, m, \quad (6)$$

где математическое ожидание E_δ находится по предельному распределению π_j^δ и выражается равенством

$$E_\delta \varphi_{t,y} = (1 - \delta) \sum_{\mu=1}^m p_{j\mu} (y^{(1)}(s_j)) r_{j\mu} (y^{(1)}(s_j)) + \lambda \delta,$$

в котором число λ меньше числа, стоящего множителем при $1 - \delta$ (оно равно среднему арифметическому одноживых прибылей, отвечающих $k - 1$ неоптимальным действиям в состоянии s_j). Это выражение, в комбинации с (5) и (6), приводит к следующему:

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{i=1}^t \varphi_i = (1 - \delta) W(y_1) + c \delta, \quad c < W(y_1)$$

с вероятностью 1. Отсюда мы видим, что при малых δ достигается цель управления — ϵ -оптимальность в сильном смысле. Теорема доказана.

Суждения о скорости сходимости $GM(n^0, \delta)$ -алгоритмов можно получить из рассмотрения моментов $E \tau^k$ случайной величины τ — немарковского момента, начиная с которого адаптивная система будет всегда находиться в δ -оптимальном режиме. К этому вопросу мы обратимся позднее.

Теперь займемся исследованием $GM(n^0, 0)$ -алгоритма, отличающегося от $GM(n^0, \delta)$ -алгоритмов тем, что

$$\lim_{n \rightarrow \infty} \delta(n) = 0.$$

Как и прежде, трансформация строк матрицы $B(t)$ происходит в моменты времени, когда на единицу возрастает величина $n^0(t) = \min_j N_j(t)$. Этот алгоритм можно tolко-ковать как предельный вариант $GM(n^0, \delta)$ -алгоритма, когда $\delta(n)$ содержит бесконечное число различных элементов и $\delta(n)$ стремится к нулю (при $n \rightarrow \infty$).

Теорема 2. $GM(n^0, 0)$ -алгоритм является асимптотически оптимальным в сильном смысле в классе $MD(S, Y)$, т. е. для любого процесса из класса и любого $\epsilon > 0$ найдется такое τ_ϵ , что при всех $t > \tau_\epsilon$

$$\frac{1}{t} \sum_1^t \varphi_j > \max_y W(y) - \epsilon.$$

Доказательство проводится по тому же плану, что и для теоремы 1. Сначала установим, что количества выборов всех правил $N_i(t)$ неограниченно растут. Если бы это было не так, то некоторое правило y_{i_0} избиралось бы конечное число раз, $N_{i_0}(t) \equiv N_{i_0}$. Отсюда следует, что $n^0(t) \equiv N_{i_0}$ при всех t . Это означает неизменность матрицы $B(t)$ начиная с некоторого момента t_0 . Пусть к моменту t_0 осуществилось n трансформаций $B(t)$, т. е. все элементы этой матрицы не меньше $\delta(n)/(k-1) > 0$, но тогда в бесконечной последовательности испытаний пра-

вило y_j непременно будет реализовано еще хотя бы один раз. Таким образом, с вероятностью единица

$$\lim_{t \rightarrow \infty} N_j(t) = \infty, \quad j = 1, \dots, k^m.$$

Начиная с момента времени τ окажется выполненным неравенство

$$V_1(t) > \max_{j \geq 2} V_j(t),$$

в прежнем предположении, что y_1 — единственное оптимальное правило. Тогда в каждом состоянии цепи будет стремиться к 1 вероятность выбора действия, сопоставляемого этому состоянию оптимальной стратегией, а вероятности остальных действий приближаться к нулю. Теперь управляемая марковская цепь $(S, \mathcal{P}_{\delta, n})$ не является однородной, но можно убедиться, что у нее существуют предельные вероятности состояний $(\pi_1(y_1), \dots, \pi_m(y_1))$. Покажем, что с вероятностью единица

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{v=1}^t \varphi_v = W(y_1),$$

означающее асимптотическую оптимальность системы, реализующей $GM(n^0, 0)$ -алгоритм. Подобно рассуждениям в предыдущей теореме, имеем

$$\frac{1}{\tau} \sum_{v=1}^t \varphi_v = \sum_{j=1}^m \frac{l_j(t)}{t} U(t), \quad (7)$$

где $l_j(t)/t \rightarrow \pi_j(y_1)$ с вероятностью 1. Для вычисления пределов последовательностей $U_j(t)$ запишем их в виде

$$\frac{1}{l_j(t)} \sum_{v=1}^{l_j(t)} \varphi_v^{(j)} = \sum_{\lambda=1}^k \frac{n_{j,\lambda}(t)}{l_j(t)} \frac{1}{n_{j,\lambda}(t)} \sum_{\mu=1}^{n_{j,\lambda}(t)} \varphi_{\mu}^{(j)}(\lambda),$$

приняв обозначения: $n_{j,\lambda}(t)$ — число, равное количеству попаданий в состояние s_j и выбору в нем действия y_λ за время t , и $(\varphi_{\mu,\lambda}^{(j)})$ — совокупность доходов, которые получены в результате переходов из состояния s_j при

действии y_{λ} . По определению $GM(n^0, 0)$ -алгоритма в каждом состоянии s_j стремится к 1 (при $t \rightarrow \infty$) частота $n_{j, \lambda_0}(t)/l_j(t)$ совершения действия y_{λ_0} , входящего в оптимальное правило y_1 , остальные частоты стремятся к 0. Легко видеть далее, что с вероятностью единица

$$\lim_{t \rightarrow \infty} \frac{1}{n_{j, \lambda_0}(t)} \sum_{\mu=1}^{n_{j, \lambda_0}(t)} \varphi_{\mu}^{(j)}(\lambda_0) = \sum_{n=1}^m p_{jn}(y_{\lambda_0}) r_{jn}(y_{\lambda_0}),$$

$$j = 1, \dots, m.$$

Таким образом, с вероятностью единица имеем при всех j

$$\lim_{t \rightarrow \infty} U_j(t) = \sum_{n=1}^m p_{jn}(y_{\lambda_0}) r_{jn}(y_{\lambda_0}).$$

Переходя на основании этого равенства к пределу в (7), получаем окончательно

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{v=0}^t \varphi_v = W(y_1).$$

Теорема доказана.

Видоизменим $GM(n^0, 0)$ -алгоритм. Новое заключается в правиле преобразования матрицы $B(t)$ вероятностей выбора действий. Условимся не менять ее в начальном периоде, пока функция $n^0(t) = \min_j V_j(t)$ не станет равной заданному числу $n_0 \geqslant 1$. После этого матрица $B(t)$ меняется на каждом такте работы алгоритма описанным ранее способом. Числа $\delta(n)$ образуют монотонно убывающую последовательность, причем

$$\lim_{n \rightarrow \infty} \delta(n) = 0, \quad \sum_{n=1}^{\infty} \delta(n) = \infty. \quad (8)$$

Условимся называть GM -алгоритм такого вида $GM(\delta)$ -алгоритмом. Его свойства как системы управления марковскими цепями с доходами сформулированы в следующей теореме.

Теорема 3. *GM (δ)-алгоритм является асимптотически оптимальным в сильном смысле в классе MD (S, Y).*

Доказательство опирается на теорему 3 из § 7 гл. III. Рассуждения, аналогичные проведенным там, позволяют утверждать, что все функции $N_j(t)$ безгранично растут, а значит, $V_j(t)$ — состоятельные оценки средних выигрышей $W(y_j)$. Значит, с некоторого момента времени в каждом состоянии вероятность выбора действия, входящего в оптимальное правило, окажется равной $1 - \delta(t)$ и будет стремиться к единице. Асимптотическая оптимальность GM (δ)-алгоритма доказывается таким же приемом, как и в теореме 2.

Итак, мы исследовали три варианта GM-алгоритмов, представляющих собой модификации алгоритмов, которые реализованы в автоматах G , применительно к управлению марковскими цепями с доходами. В рассмотренных алгоритмах находятся не характеристики эволюции цепей, заключенные в матрицах вероятностей переходов, а результаты использования правил — оценки средних выигрышней, т. е. косвенные характеристики управляемого процесса. Теперь мы обратимся к алгоритмам, требующим сопирания информации о процессе, т. е. в нашем случае — оценки вероятностей переходов из состояния в состояние под действием каждого управления $y \in G$. Сверх того, предполагается оценивание всех матриц доходов. Знание перечисленных матриц позволяет вычислить оптимальную стратегию. Опишем один из методов ее вычисления, именуемый *итеративным методом*. Матрицы $\mathcal{T}^{(y)}$ и $R^{(y)}$ известны для всех $y \in Y$.

Выберем некоторое правило $y_0 = (y_0^{(1)}, \dots, y_0^{(m)})$, где $y_0^{(i)} = \sigma_0(s_i)$, и вычислим величины одн шаговых прибылей для каждого состояния s_j и отвечающего ему действия $y_0^{(j)}$

$$q_j = q(j, y_0^{(j)}) = \sum_{l=1}^m p_{jl}(y_0^{(j)}) r_{jl}(y_0^{(j)}), \quad j = 1, \dots, m.$$

Теперь образуем систему уравнений m -го порядка (δ_{ij} — символ Кронекера)

$$\sum_{j=1}^m (p_{ij}(y_0^i) - \delta_{ij}) v_j - W = -q_i, \quad i = 1, \dots, m, \quad (9)$$

в которой неизвестны W и m «весов» v_1, \dots, v_m . Условимся здесь и далее принимать $v_m=0$. Тогда из приведенной системы можно найти m неизвестных величин. Это решение является первым шагом одного итерационного цикла, второй шаг его заключается в отыскании для каждого состояния s_i такого действия $y^{(i)}$, которое максимизирует величину

$$q_i(y) + \sum_{j=1}^m p_{ij}(y) v_j, \quad i = 1, \dots, m,$$

с найденными на предыдущем шаге весами. Полученный набор действий $y_1=(y_1^{(1)}, \dots, y_1^{(m)})$ служит новым правилом, его отыскание завершает итерационный цикл. Когда на первом шаге следующего цикла решают систему уравнений вида (9), то величина W означает предельный средний выигрыш в единицу времени при использовании стратегии y_1 , т. е. $W=W(y_1)$. Одновременно находят веса (снова положив $v_m=0$) и т. д.

Доказывается, что эта процедура заканчивается после конечного числа циклов вычислением оптимального правила y_{opt} и отвечающего ему предельного выигрыша $W(y_{\text{opt}})=\max_y W(y)$.

Перейдем к описанию алгоритмов, которые мы назовем *RZ-алгоритмами*. Управлению подвергаются снова эргодические марковские цепи с доходами из класса $MD(S, Y)$.

В каждый момент времени подсчитывается одна из km^2 функций $n_{ij}^{(l)}(t)$, означающих число переходов за время t из состояния s_i в состояние s_j под действием y_l . С помощью этих функций оцениваются элементы неизвестных заранее матриц $\mathcal{P}^{(y)}$ по обычным формулам

$$\hat{p}_{ij}^{(l)}(t) = \begin{cases} \frac{n_{ij}^{(l)}(t)}{n_i^{(l)}(t)}, & \text{если } n_i^{(l)}(t) = \sum_{\mu=1}^m n_{i\mu}^{(l)}(t) > 0, \\ 0 & \text{в противном случае.} \end{cases}$$

Матрицы $R^{(y)}$ находятся точно. Элементами их служат доходы $r_{ij}^{(l)}$, наблюдаемые в моменты соответствующих переходов. Начальные значения матриц $\hat{R}(y)$, являю-

щихся оценками $R^{(y)}$, полагаем равными 0. Поэтому в случае, если какой-нибудь переход не наступил, соответствующий элемент матрицы остается нулевым. Всего RZ -алгоритмы требуют хранения $3km^2$ чисел.

Действия $y \in Y$ выбираются с помощью распределений вероятностей на Y , сопоставленных каждому состоянию цепи. Они записываются в виде матрицы порядка $m \times k$

$$B(t) = \|B_{ij}(t)\|,$$

номерами строк которой служат номера состояний цепи, а номерами столбцов номера действий.

В начальный момент все элементы матрицы одинаковы, $B_{ij}(0)=1/k$, и они не меняются до тех пор, пока хотя бы одна величина $n_i(t) = \min_l n_i^{(l)}(t)$ остается равной 0. После того как впервые выполнится условие $\min_i n_i(t) = 1$, по матрицам оценок $\hat{P}^{(y)}$, $\hat{R}^{(y)}$, $y \in Y$, с помощью итеративного метода в каждый момент вычисляются «псевдооптимальные» правила $\tilde{y} = (y^{(1)}, \dots, y^{(m)})$. Состоянию s_i цепи мы сопоставляем «лучшее» действие $y^{(i)}$, расположенное на i -й компоненте правил. Если оказывается, что итеративный метод дает несколько таких правил, мы выбираем одно из них. Способ изменения вероятностей выбора действий определяется функцией $\delta(n)$, удовлетворяющей следующим условиям:

$$0 < \delta(n) < \frac{1}{2}, \quad \lim_{n \rightarrow \infty} \delta(n) = \delta \geq 0,$$

причем стремление к 0 — монотонное.

В каждый момент t меняется ровно одна строка матрицы $B(t)$, номер которой равен номеру текущего состояния цепи. Преобразование осуществляется по формуле

$$B_{ij}(t) = \begin{cases} 1 - \delta(n_i(t)), & y_j = \tilde{y}_t(s_i), \\ \frac{\delta(n_i(t))}{k-1}, & y_j \neq \tilde{y}_t(s_i), \end{cases}$$

где $\tilde{y}_t(s_i)$ — значение в состоянии s_i оптимального либо «псевдооптимального» правила, найденного в этот момент времени с помощью итеративного метода.

Если последовательность $\delta(n)$ сходится к положительному пределу δ , будем предполагать, так же как в случае *GM*-алгоритмов, что эта последовательность содержит конечное число членов. Значит, после конечного количества преобразований матрицы $B(t)$ система управлений выйдет на δ -режим, но он может быть сначала не δ -оптимальным из-за неточности в исходных данных для итеративного метода вычисления оптимального правила. Однако справедлив следующий факт.

Теорема 4. *RZ-алгоритмы, в которых $\delta > 0$, представляют собой ϵ -оптимальные в сильном смысле семейства адаптивных систем. Если $\delta=0$, то такие RZ-алгоритмы асимптотически оптимальны в сильном смысле.*

Доказательство проводится по тому же плану, что и доказательства теорем 1—3. Прежде всего убеждаемся, что для каждой пары состояний s_i, s_j , для которой $p_{ij}(y_l) > 0$ при некотором l , имеет место $n_{ij}^{(l)}(t) \rightarrow \infty$. Действительно, согласно правилу изменения вероятностей выбора действий гарантируется, что в каждом состоянии s_i любое действие будет приложено бесконечное число раз. Это означает, что все переходы $s_i \rightarrow s_j$ (под действием y_l), имеющие положительную вероятность, непременно осуществляются бесконечно много раз. Следовательно, оценки $\hat{p}_{ij}^{(l)}(t)$ состоятельные, т. е. для всех троек индексов (i, j, l)

$$P(\lim_{t \rightarrow \infty} \hat{q}_{ij}^{(l)}(t) = p_{ij}(y_l)) = 1.$$

Поэтому все матрицы $\hat{\mathcal{P}}^{(y)}$ сходятся к истинным матрицам вероятностей переходов $\mathcal{P}^{(y)}$. Ранее было указано, что оценки матриц доходов $\hat{R}^{(y)}$, начиная с некоторого момента становятся равными матрицам доходов с точностью до тех элементов, которым отвечают в соответствующей матрице $\mathcal{P}^{(y)}$ нулевые вероятности переходов. Из сказанного вытекает, что вычисленная с помощью итеративного метода «псевдооптимальное» правило оказывается оптимальным по истечении достаточно большого времени, когда оценки матриц $\hat{\mathcal{P}}^{(y)}$ оказываются близкими к истинным значениям. Проверка свойств ϵ -оптимальности или асимптотической оптимальности, в зависимости от стремления $\delta(n)$ к положительному числу δ или к нулю, про-

водится такими же способами, как и в предыдущих теоремах. На этих рассуждениях мы не останавливаемся.

Теперь мы обратимся к уже поставленному ранее вопросу о существовании (конечности) моментов случайной величины τ — немарковского момента, начиная с которого алгоритмы начинают осуществлять цель управления (оказываются в δ -оптимальном режиме (если $\delta > 0$) или же вероятности оптимальных действий во всех состояниях (в случае $\lim_{n \rightarrow \infty} \delta(n) = 0$) устремляются к 1).

Сформулируем определения для обоих типов алгоритмов.

Временем обучения GM-алгоритма называется случайный (немарковский) момент времени τ такой, что при $t < \tau$

$$\max_{j \leq L} V_j(t) \leq \max_{i > L} V_i(t),$$

где $y_1, \dots, y_L, L < k^m$ — оптимальные правила, а при всех $t \geq \tau$

$$\max_{j \leq L} V_j(t) > \max_{i > L} V_i(t).$$

Согласно этому определению до момента τ допускается выполнение неравенства $\max_{j \leq L} V_j(t) > \max_{i > L} V_i(t)$, но затем оно может нарушаться.

Временем обучения RZ-алгоритма называется случайный момент τ такой, что найденное по оценкам матриц $\hat{\mathcal{P}}^{(y)}$ с помощью итерационного метода «псевдооптимальное» правило не совпадает в моменты $t < \tau$ с оптимальным правилом, а при всех $t \geq \tau$ оно равно одному из них.

Главное для нас свойство случайной величины τ сформулируем сразу для GM- и RZ-алгоритмов.

Теорема 5. Все моменты времени обучения τ конечны, если выполнено одно из двух условий: а) при некотором $\lambda \in (0, 1/2)$

$$\overline{\lim}_{t \rightarrow \infty} n^\lambda \delta(n) \leq c < \infty,$$

б) конечная последовательность $\delta(n)$ сходится к числу $\delta > 0$.

Доказательство мы проведем лишь для $GM(n^0, \delta)$ -алгоритма. В остальных вариантах требуются незначительные изменения рассуждений.

Рассматриваемые нами моменты изображаются формулой

$$E\tau^l = \sum_{t=1}^{\infty} t^l P(\tau=t).$$

Поэтому теорема будет доказана, если удастся получить оценку, например, вида $P(\tau=t) < ce^{-\alpha t}$ (α, c — положительные числа). Неравенство такого вида мы установим для простоты в предположении, что оптимальное управление единственно — y_1 . В силу соотношений

$$\begin{aligned} P(\tau=t) &= P(V_1(t) \leqslant \\ &\leqslant \max_{i \geqslant 2} V_i(t), V_1(t+s) > \max_{i \geqslant 2} V_i(t+s) \text{ при всех } s \geqslant 1) \leqslant \\ &\leqslant P(V_1(t) \leqslant \max_{i \geqslant 2} V_i(t)) \end{aligned}$$

остается доказать, что

$$P(V_1(t) \leqslant \max_{i \geqslant 2} V_i(t)) < ce^{-\alpha t}$$

при всех достаточно больших t . Имеем

$$\begin{aligned} P(V_1(t) \leqslant \max_{i \geqslant 2} V_i(t)) &\leqslant \sum_{i=2}^{k^m} P(V_1(t) \leqslant V_i(t)) \leqslant \\ &\leqslant (k^m - 1) \max_{i \geqslant 2} P(V_1(t) < V_i(t)). \end{aligned}$$

Обозначим

$$\varepsilon_i = W(y_1) - W(y_i), \quad \varepsilon = \min_{i \geqslant 2} \varepsilon_i.$$

Тогда

$$\begin{aligned} P(V_1(t) \leqslant \max_{i \geqslant 2} V_i(t)) &\leqslant \\ &\leqslant P\left(|V_1(t) - W(y_1)| > \frac{\varepsilon_i}{3}\right) + P\left(|V_i(t) - W(y_i)| > \frac{\varepsilon_i}{3}\right) \leqslant \\ &\leqslant 2 \max_{i \geqslant 1} P\left(|V_i(t) - W(y_i)| > \frac{\varepsilon_i}{3}\right). \end{aligned}$$

Вероятность в правой части оценивается

$$\begin{aligned} \mathsf{P}\left(|V_i(t) - W(y_i)| > \frac{\epsilon}{3}\right) &= \\ &= \sum_{n=0}^t \mathsf{P}\left(|V_i(t) - W(y_i)| > \frac{\epsilon}{3} \mid N_i(t) = n\right) \mathsf{P}(N_i(t) = n) \leqslant \\ &\leqslant \sum_{n=0}^t \mathsf{P}\left(\left|\frac{1}{n} \sum_{v=1}^n \varphi_{t,v} - W(y_i)\right| > \frac{\epsilon}{3}\right) \mathsf{P}(N_i(t) = n), \end{aligned}$$

причем среднее арифметическое в скобках полагаем равным 0 при $n=0$. Воспользуемся теперь следующей леммой.

Л е м м а. Пусть (S, \mathcal{F}) — эргодическая марковская цепь с t состояниями и предельным распределением $(\pi_1, \dots, \dots, \pi_m)$. Рассмотрим функционал $\varphi_t = \varphi(s_t)$ на состояниях цепи и обозначим

$$C_t = \sum_{v=1}^t \varphi(s_v), \quad W = \sum_{i=1}^m \varphi(s_i) \pi_i.$$

При $x \geqslant x_0 > 0$ существуют такие положительные постоянные $c=c(x_0)$ и $\beta=\beta(x_0)$, что при всех $t \geqslant 1$

$$\mathsf{P}\left(\left|\frac{C_t}{t} - W\right| > x\right) \leqslant ce^{-\beta t}.$$

Применим этот результат *) к оценке вероятности

$$\mathsf{P}\left(\left|\frac{1}{n} \sum_{v=0}^n \varphi_{t,v} - W(y_i)\right| > \frac{\epsilon}{3}\right) \leqslant c_i e^{-\beta_i t}, \quad i = 1, \dots, k^m.$$

*) Наметим доказательство леммы. В ее предпосылках справедлива центральная предельная теорема: при некотором $B > 0$

$$\lim_{t \rightarrow \infty} \mathsf{P}\left(\left|\frac{C_t - Wt}{\sqrt{Bt}}\right| > x\right) = \sqrt{\frac{2}{\pi}} \int_x^{\infty} e^{-z^2/2} dz, \quad x > 0.$$

(Продолжение сноски см. на стр. 256.)

Положив теперь $\beta = \min_i \beta_i$, имеем (h выберем ниже)

$$\mathbb{P}\left(|V_i(t) - W(y_i)| > \frac{\varepsilon}{3}\right) \leq$$

$$\leq c_i \sum_{n=0}^t e^{-\beta n} \mathbb{P}(N_i(t) = n) = \sum_{n=0}^h + \sum_{n=h+1}^t.$$

Вторая сумма справа оценивается сразу,

$$\begin{aligned} \sum_{n=h+1}^t e^{-\beta n} \mathbb{P}(N_i(t) = n) &< \sum_{n=h+1}^t e^{-\beta n} = \\ &= e^{-\beta(h+1)} t \sum_{n=0}^{h+1} e^{-\beta n} = e^{-\beta(h+1)} \frac{1 - e^{-\beta(t-h)}}{1 - e^{-\beta}} < b e^{-\beta(h+1)}, \end{aligned}$$

а для оценки первой суммы введем вспомогательную последовательность случайных величин

$$\eta_i = \begin{cases} 1 & \text{с вероятностью } \frac{\delta}{k-1}, \\ 0 & \text{с вероятностью } 1 - \frac{\delta}{k-1}, \end{cases}$$

причем $E\eta_i = \frac{\delta}{k-1}$. Составим из этих величин сумму $\tilde{N}_i(t) = \sum_{l=1}^t \eta_l$, тогда $E\tilde{N}_i(t) = \frac{\delta}{k-1} t$. Полагая $h = \frac{\delta}{2(k-1)} t$,

Согласно известным оценкам

$$\int_x^\infty e^{-z^2/2} dz \leq \frac{1}{\sqrt{x}} e^{-x^2/2}$$

отсюда получаем

$$\begin{aligned} \mathbb{P}\left(\left|\frac{C_t}{t} - W\right| > x\right) &= \mathbb{P}\left(\left|\frac{C_t - Wt}{\sqrt{Bt}}\right| > x \sqrt{\frac{t}{B}}\right) \leq \\ &\leq ce^{-\frac{x^2}{2B} t} \leq ce^{\frac{x_0^2}{2B} t} = ce^{-\beta t} \end{aligned}$$

(здесь $\beta = \frac{x_0^2}{2B}$).

приходим к цепочке неравенств, последнее из которых получено с помощью леммы 1 из § 5 гл. III,

$$\begin{aligned} \mathsf{P}\left(N_i(t) \leq \frac{\delta}{2(k-1)} t\right) &\leq \mathsf{P}\left(\tilde{N}_i(t) \leq \frac{\delta}{2(k-1)} t\right) = \\ &= \mathsf{P}\left(\tilde{N}_i(t) - \mathsf{E}\tilde{N}_i(t) \leq -\frac{\delta}{2(k-1)} t\right) \leq e^{-\frac{\delta^2}{2(k-1)^2} t}. \end{aligned}$$

Объединяя полученные оценки, приходим к следующей:

$$\begin{aligned} \mathsf{P}\left(|V_i(t) - W(y_i)| > \frac{\epsilon}{3}\right) &\leq \\ &\leq c_i b e^{-\frac{\beta\delta}{2(k-1)} t} + e^{-\frac{\delta^2}{2(k-1)^2} t} \leq d e^{-\alpha t}, \end{aligned}$$

где $d = \max_i c_i b + 1$, $\alpha = \min\left(\frac{\beta\delta}{2(k-1)}, \frac{\delta^2}{2(k-1)^2}\right)$. Обозначив $c = 2d(k^m - 1)$, приходим к искомому результату

$$\mathsf{P}(\tau = t) \leq \mathsf{P}(V_1(t) \leq \max_{i \geq 2} V_i(t)) < ce^{-\alpha t}.$$

Тем самым теорема доказана для $\delta > 0$. Нетрудно усмотреть видоизменение доказательства в случаях, когда $\delta(n) \rightarrow 0$.

Опираясь на теорему 5, можно дать верхние оценки моментов времени обучения, в частности среднего времени обучения. Это делается аналогично рассуждениям § 5 гл. III. Моменты времени обучения лишь косвенно характеризуют скорость сходимости алгоритмов. Непосредственные суждения о скорости делаются из оценок близости суммарного выигрыша V_t к величине \bar{W} в той или иной метрике. Например, для некоторых вариантов асимптотически оптимальных алгоритмов удается установить оценку вида

$$\mathsf{E}|\bar{W} - V_t|^2 = O(t^{-1/2} + \delta(t)).$$

Приближенный характер этих оценок не дает пока оснований для надежных суждений о скорости сходимости алгоритмов и тем более об «оптимальном» (по быстроте). Подобного рода сведения можно получить имитацией адаптивных систем на ЦВМ, т. е. экспериментальным

путем. Поэтому ниже мы ограничиваемся сопоставлением перечисленных алгоритмов лишь по сложности их реализации.

GM-алгоритмы требуют хранения $2k^m$ чисел, входящих в совокупности $(N_i(t))$ и $(V_i(t))$, а *RZ*-алгоритмы $3km^2$ чисел из $(n_{ij}^{(l)}(t))$, $(\tilde{p}_{ij}^{(l)}(t))$ и $(\tilde{r}_{ij}^{(l)}(t))$. Кроме того, оба алгоритма нуждаются в mk числах матрицы $B(t)$. В типических случаях *GM*-алгоритмы предъявляют более тяжелые требования к необходимому объему памяти, чем *RZ*-алгоритмы. Количество операций на каждом такте работы алгоритмов также не одинаковы. Помимо приблизительно одинаковых затрат на преобразование матрицы $B(t)$, эти алгоритмы ведут вспомогательные вычисления. У *GM*-алгоритмов это добавление единиц к k^{m-1} числам $N_i(t)$ и пересчет такого же числа эмпирических средних выигрышей $V_i(t)$. *RZ*-алгоритмы пересчитывают строку матрицы $\hat{\mathcal{P}}^{(y)}$ и, возможно, $R^{(y)}$, а затем вычисляют по имеющимся сведениям очередное «псевдо-оптимальное» правило (спустя время обучения τ оно уже оказывается оптимальным). Использование соответствующих вычислительных — итеративных — методов требует значительного числа операций, и для управления процессом в реальном времени могут оказаться необходимыми вычислительные средства с высоким быстродействием.

§ 4. Задачи с целевыми неравенствами

Рассматриваются объекты управления, которые описываются разностными уравнениями конечного порядка $l \geqslant 1$

$$\xi_t = g(\xi_{t-l}, y_{t-l}^{t-1}), \quad t \geqslant l,$$

с надлежащими начальными условиями. Классы таких объектов выделяются предположениями относительно функций $g(\cdot, \cdot)$. Согласно принятой нами терминологии эти объекты относятся к управляемым l -связным детерминированным марковским процессам. Мы не будем здесь непременно отождествлять управляемый процесс ξ_t с наблюдаемым. Последний (η_t) изображается зависимостью $\eta_t = h(\xi_t, y^{t-1})$.

Управление объектом осуществляется посредством правил выбора действий $f(\eta_{t-l}^{t-1}, y_{t-l}^{t-1})$, принадлежащих некоторому множеству D допустимых правил, причем l' не обязательно совпадает с l .

Примем фазовое пространство X и пространство управлений Y вещественными гильбертовыми.

В качестве цели управления для охарактеризованного класса процессов потребуем во все моменты времени, начиная с некоторого, выполнение неравенства

$$\varphi_t = \varphi(\xi^t, y^t) > 0$$

для заданного на траекториях процесса функционала φ . Впредь всегда подразумеваем, что значения функционала наблюдаемы.

Адаптивная постановка указанной задачи управления означает, как обычно, что неизвестен точный вид уравнения, определяющего процесс. Следовательно, речь идет об отыскании стратегии, приводящей к цели на классе процессов.

Далее излагаются идеи и методология одного подхода к синтезу адаптивных систем, которые за конечное время обеспечивают выполнение неравенств $\varphi_t > 0$ для каждого процесса класса.

Пусть для любого процесса из класса можно построить оптимальное правило выбора управлений, приводящее к цели $y_t = f(\eta^t, \theta)$. В него входит векторный параметр θ , его значение зависит от конкретного процесса из класса. Функциональный вид f предполагается известным, но недостаток сведений об управляемом процессе не позволяет подставить в эту функцию значение параметра θ . Отметим в качестве примера класс L_X процессов, описываемых линейными уравнениями

$$\xi_t = \sum_{i=1}^l (A_i \xi_{t-i} + B_i y_{t-i}) + \zeta(t).$$

Во многих интересных случаях оптимальное управление для них имеет вид

$$y_t = \sum_{j=1}^N \theta_j q_j(\eta_t),$$

где q_j — известные функции наблюдаемого процесса, одинаковые для всех процессов класса L_x , коэффициенты $(\theta_1, \dots, \theta_N) = \theta$ зависят от конкретного процесса этого класса. Существование указанного представления оптимального управления принципиально важно для дальнейшего. При отсутствии сведений о всех параметрах управляемого процесса оптимальное управление надо искать в форме

$$g_t = f(\eta^t, \theta_t),$$

где θ_t должна вычисляться каким-нибудь способом по результатам наблюдений за эволюцией процесса и в зависимости от выполнения (или невыполнения) неравенства $\varphi_t > 0$. Этот замысел реализован в излагаемом ниже подходе.

Синтез требуемой адаптивной системы управления означает построение алгоритма решения за конечное число шагов системы неравенств, вообще говоря, бесконечной. К такой системе мы приходим следующим образом: в момент t для функционала $\varphi_t = \varphi(\xi^t, y^t)$ имеем

$$\varphi(\xi^t, y^{t-1}, y_t) = \varphi(\xi^t, y^{t-1}, f(\eta^t, \theta_t)) = \varphi_t(\theta_t).$$

Нахождение таких θ_t , для которых $\varphi_t(\theta_t) > 0$, обеспечивает выполнение целевого неравенства и указывает одновременно очередное действие. Отметим еще один вариант задачи для функционалов $\varphi_t = \varphi(\xi^t)$. Интересующие нас неравенства получаются в этом случае так:

$$\begin{aligned} \varphi_{t+1} &= \varphi(\xi^t, \xi_{t+1}) = \varphi(\xi^t, g(\xi_{t-l'+1}^t, y_{t-l'+1}^t)) = \\ &= \varphi(\xi^t, g(\xi_{t-l'+1}^t, y_{t-l'+1}^{t-1}, y_t)) = \\ &= \varphi(\xi^t, g(\xi_{t-l+1}^t, y_{t-l+1}^{t-1}, f(\eta^t, \theta_t))) = \varphi_{t+1}(\theta_t). \end{aligned}$$

Здесь значения функционала регистрируются на следующем такте после фиксации параметра θ_t .

Важнейшей особенностью редуцированной таким образом постановки задачи является то, что в каждый момент времени t известно значение (либо знак) лишь функционала φ_t , но не будущие его значения, которые зависят от будущего выбора параметров $\theta_{t+1}, \theta_{t+2}, \dots$. Эти последние выбираются после «предъявления» соответствующих

значений функционалов. Иными словами, требуется за конечное число шагов решить счетную систему «не показанных» неравенств. Ясно, что сформулированную таким образом задачу возможно решить не всегда. Одному общему результату мы предпоследнему несколько определений, а затем примеров.

Интересующая нас система неравенств имеет вид

$$\varphi(\xi_t, \theta_t) > 0, \quad t \geq 1.$$

Алгоритм решения этой системы называется *конечно-сходящимся* (кратко КСА), если существует конечное t^0 такое, что при всех $t > t^0$ и любом ξ из допустимой области фазового пространства X выполнено $\varphi(\xi, \theta_t) > 0$. Число моментов времени, в которые $\varphi \leq 0$, называют *числом исправлений (коррекций)* алгоритма.

Среди КСА особый интерес вызывают (благодаря простоте реализации) рекуррентные процедуры. Продемонстрируем высказанные общие положения несколькими примерами КСА.

Предложены линейные однородные неравенства $\varphi(x, \theta) = (x, \theta) > 0$, здесь (x, θ) — скалярное произведение элементов гильбертова пространства X . Пусть система неравенств порождается счетным множеством точек $\{x_t\}$ из сферы радиуса r , $\|x_t\| \leq r$. Допустим, что существуют число $\varepsilon_* > 0$ и элемент θ_* такие, что $(\theta_*, x_t) > \varepsilon_*$ при всех x_t . Следующая процедура является конечно-сходящимся алгоритмом решения счетной системы неравенств $(\theta, x_t) > 0$:

$$\theta_{t+1} = \begin{cases} \theta_t, & \text{если } (\theta_t, x_t) > 0, \\ \theta_t + \left[a_t - b_t \frac{(\theta_t, x_t)}{(x_t, x_t)} \right] x_t, & \text{если } (\theta_t, x_t) \leq 0. \end{cases}$$

Здесь $0 < a' < a_t < a''$, $0 \leq b_t \leq 2$ при $t \geq 1$. Можно указать верхнюю оценку числа исправлений алгоритма.

Теперь рассмотрим систему неравенств типа «полоска»

$$\varphi_t = |(\theta, x_t) + \lambda(t)| < h,$$

причем $|x_t| \leq r$ и существуют такие θ_* и $h_* \in (0, h/2)$, что $|(\theta_*, x_t) + \lambda(t)| \leq h_* < h/2$ при всех $t \geq 1$. Для нее

конечно-сходящимся алгоритмом служит процедура

$$\theta_{t+1} = \begin{cases} \theta_t, & \text{если } \varphi_t < h, \\ \theta_t - \frac{\varphi_t}{(x_t, x_t)} x_t, & \text{если } \varphi_t \geq h. \end{cases}$$

Наконец, обратимся к системе квадратичных неравенств

$$\varphi_t(\theta) = (K_t(\theta - a_t), (\theta - a_t)) < \alpha_t, \quad \alpha_t > 0,$$

где K_t — симметрические матрицы такие, что при всяком $x \in X$

$$(Hx, x) \geq (K_t x, x) \geq 0;$$

H — симметрическая положительно определенная матрица. Кроме того, $\alpha_t \geq \alpha > 0$. Решение этой системы доставляет следующий КСА:

$$\theta_{t+1} = \begin{cases} \theta_t, & \text{если } \varphi_t(\theta_t) < \alpha_t, \\ \theta_t - b_t H^{-1} K_t(\theta_t - a_t), & \text{если } \varphi_t(\theta_t) \geq \alpha_t, \end{cases}$$

где $0 < b' \leq b_t \leq b'' < 2(1-\rho)$, $\rho > 0$. В этом, как и в предыдущем, случае удается дать верхнюю оценку числа исправлений алгоритма.

Приведем достаточные условия существования КСА в виде рекуррентной процедуры $\theta_{t+1} = v_t(x_t, \theta_t)$. Мы предполагаем, что элементы последовательности x_t принадлежат ограниченному подмножеству гильбертова пространства X , и не вдаемся в способ получения этой последовательности. Величины θ — элементы гильбертова пространства Θ со скалярным произведением (θ', θ'') и нормой $\|\theta\| = \sqrt{(\theta, \theta)}$.

Теорема 1. Пусть функция $\varphi(x, \theta)$ подчинена следующим условиям:

а) она определена на $X_0 \times \Theta_0$, где $X_0 \subseteq X$ — ограниченное множество, а $\Theta_0 \subseteq \Theta$ выпуклое;

б) она дифференцируема по θ и производная $\nabla_\theta \varphi(x, \theta)$ ограничена на $X_0 \times \Theta_0 \cap \{\varphi(x, \theta) \leq 0\}$;

в) существует элемент $\theta_* \in \Theta_0$ такой, что для всех $x \in X_0$

$$\varphi(x, \theta_*) \geq \varepsilon_* > 0;$$

г) для любых $(x, \theta) \in X_0 \times \Theta_0 \cap \{\varphi(x, \theta) \leq 0\}$ имеет место

$$(\nabla_\theta \varphi(x, \theta), \theta_* - \theta) \geq \varphi(x, \theta_*) - \varphi(x, \theta).$$

Пусть далее, $\tilde{X} = \{x_t\}$ — счетная последовательность из X_0 и $\theta_1 \in \Theta_0$ произвольно.

Зададим рекуррентную процедуру $\theta_{t+1} = v_t(x_t, \theta_t)$, где

$$v_t(x_t, \theta_t) =$$

$$= \begin{cases} \theta_t, & \text{если } \varphi(x_t, \theta_t) > 0, \\ P[\theta_t - a(t) \nabla_\theta \varphi(x_t, \theta_t)], & \text{если } \varphi(x_t, \theta_t) \leq 0; \end{cases} \quad (1)$$

P — оператор проектирования на множество *) Θ_0 ,

$$a(t) = \frac{1}{n(t)} - b(t) \frac{\varphi(x_t, \theta_t)}{\|\nabla_\theta \varphi(x_t, \theta_t)\|^2},$$

$b(t) \in [0, 2]$, $n(t)$ равно числу исправлений параметра θ_t за время t , т. е. (с учетом $n(1)=1$) определяется равенством

$$n(t+1) = \begin{cases} n(t), & \text{если } \varphi(x_t, \theta_t) > 0, \\ n(t) + 1, & \text{если } \varphi(x_t, \theta_t) \leq 0. \end{cases}$$

Эта процедура представляет собой конечно-ходящийся алгоритм решения системы неравенств $\varphi(x_t, \theta_t) > 0$ для произвольного счетного множества X .

Доказательство. Покажем, что θ_t стремится (и достаточно быстро) к θ_* . Обозначая через χ_t функцию

$$\chi_t = \begin{cases} 1, & \text{если } \varphi(x_t, \theta_t) \leq 0, \\ 0, & \text{если } \varphi(x_t, \theta_t) > 0, \end{cases}$$

в силу алгоритма (1) и условий теоремы получим

$$\begin{aligned} \|\theta_{t+1} - \theta_*\|^2 &= \|P[\theta_t - \chi_t a(t) \nabla_\theta \varphi(x_t, \theta_t)] - \theta_*\|^2 \geq \\ &\geq \|\theta_t - \theta_* - \chi_t a(t) \nabla_\theta \varphi(x_t, \theta_t)\|^2 = \\ &= \|\theta_t - \theta_*\|^2 - 2\chi_t a(t) (\nabla_\theta \varphi(x_t, \theta_t), \theta_t - \theta_*) + \\ &+ \chi_t a^2(t) \|\nabla_\theta \varphi(x_t, \theta_t)\|^2 \geq \|\theta_t - \theta_*\|^2 + \chi_t [2a(t) \varepsilon_* - \varphi(x_t, \theta_t) - \\ &- a^2(t) \|\nabla_\theta \varphi(x_t, \theta_t)\|^2]. \end{aligned}$$

*) Это означает, что результат действия оператора P удовлетворяет соотношению

$$\|P\theta - \theta\| = \min_{\theta' \in \Theta_0} \|\theta' - \theta\|.$$

Учитывая вид функции $a(t)$, с помощью элементарных выкладок убеждаемся, что при больших $n(t)$ справедливо неравенство

$$\chi_t [2a(t)(\varepsilon_* - \varphi(x_t, \theta_t)) + a^2(t) \|\nabla_{\theta} \varphi(x_t, \theta_t)\|^2] \leq \frac{\varepsilon_*}{n(t)} \chi_t.$$

Это приводит к оценке

$$\|\theta_t - \theta_*\|^2 - \|\theta_{t+1} - \theta_*\|^2 \geq \frac{\varepsilon_*}{n(t)} \chi_t.$$

Суммируя эти неравенства, получим

$$\|\theta_1 - \theta_*\|^2 - \|\theta_{t+1} - \theta_*\|^2 \geq \varepsilon_* \sum_{s=1}^t \frac{\chi_s}{n(s)} = \varepsilon_* \sum_{s=1}^{n(t)} \frac{1}{s},$$

откуда непосредственно следует конечность предела $\lim_{t \rightarrow \infty} n(t)$, что и утверждалось.

Покажем теперь, как воспользоваться приведенными результатами для синтеза адаптивных систем, предназначенных для управляемых динамических систем, описываемых разностными уравнениями $\xi_t = g(\xi_{t-1}^{t-1}, y_{t-1}^{t-1})$ при (быть может) ограничениях на управления и фазовые координаты. Сформулирована цель: удовлетворить неравенствам $\varphi(\xi^t, y^t) > 0$, причем известна структура обеспечивающего эту цель правила $f(\eta^t, \theta)$ (наблюдаемый процесс отличен от управляемого). Искомая адаптивная система строится так: находится конечно-сходящийся алгоритм решения счетной системы неравенств

$$\varphi(\xi^t, y^{t-1}, f(\eta^t, \theta_t)) > 0, \quad t \geq 1.$$

Получаемые в процессе решения неравенств значения параметра θ_t используют для вычисления управляемых воздействий $y_t = f(\eta^t, \theta_t)$.

Развитые выше соображения мы применим к задаче адаптивной стабилизации решений линейных разностных уравнений.

Задан класс L процессов, представляющих собой решения уравнений

$$\xi_t = A_1 \xi_{t-1} + \dots + A_t \xi_{t-t} + B_1 y_{t-1} + \dots + B_t y_{t-t} + \zeta(t), \quad (2)$$

где ξ_t и y_t — элементы m -мерного евклидова пространства R_m с обычной метрикой, A_i , B_i — ($m \times m$)-матрицы *) и $\zeta(t)$ — возмущение («внешняя сила»). Предполагается, что структура уравнения (2) известна, но матрицы A_i , B_i и возмущение ζ неизвестны. Управления y_t находятся в нашем распоряжении, но стеснены ограничением (при всех t)

$$\|y_t\| \leq Q.$$

Фазовые координаты ξ_t наблюдаемы и удовлетворяют условию

$$\|\xi_t\| \leq R.$$

Числа Q и R заданы.

Требуется синтезировать адаптивную систему, которая спустя конечное время после начала функционирования приведет к обеспечению следующей цели управления: должно выполняться неравенство

$$\|\xi_t\| < r, \quad r < R,$$

при всех $t > t^0$.

Определенный выше класс процессов L чрезмерно широк, и назначенная цель в нем не достигается. Действительно, поскольку свойства возмущения $\zeta(t)$ заранее неизвестны и оно не наблюдаемо, то в случаях, когда $|\zeta(t)| > r$, выполнение цели гарантировать нельзя. Сверх того, необходимы ограничения на коэффициенты уравнения (2). Введем соответствующее определение.

Рассмотрим разностное уравнение

$$B_1 y_t + B_2 y_{t-1} + \dots + B_l y_{t-l+1} = z_t, \quad (3)$$

коэффициенты B_i , которого взяты из уравнения (2), и составим характеристический многочлен $\beta(\lambda) = \det(B_1 \lambda^l + B_2 \lambda^{l-1} + \dots + B_{l-1} \lambda + B_l)$. Уравнение (3) называется диссипативным, если все корни многочлена $\beta(\lambda)$ лежат внутри единичного круга. Нам понадобится следующая простая лемма.

*) Нормой матрицы $M = (m_{ij})$ назовем число $\|M\| = \sqrt{\sum_{i,j} m_{ij}^2}$.

Л е м м а. *Решение диссипативного уравнения (3) обладает свойством: существует $\nu > 0$ такое, что для любых t и $c > 0$ из неравенства*

$$\|\boldsymbol{z}_{t+m}\| < c, \quad m \geq 0,$$

и из $\boldsymbol{y}_{t-1} = \boldsymbol{y}_{t-2} = \dots = \boldsymbol{y}_{t-l+1} = 0$ следует

$$\|\boldsymbol{y}_{t+m}\| \leq \nu c.$$

В этом утверждении величина ν зависит от матриц B_1, \dots, B_l . Ниже, без дополнительных пояснений, считаем ν наименьшей из таких величин.

Теперь мы рассматриваем класс \tilde{L} процессов такой, что $\tilde{L} \subset L$ и выполнены дополнительные предположения:
1) возмущения ограничены: $\|\xi(t)\| < r$ при всех t ;
2) корни уравнения $\beta(\lambda)=0$ лежат внутри единичного круга. Течение каждого процесса происходит на системе конечных неперекрывающихся временных интервалов $\Delta_1, \Delta_2, \dots$ длиной выше l тактов каждый. В первые $l-1$ моменты времени интервала Δ_i назначаются начальные значения процесса $\xi_t (t=t_i, t_i+1, \dots, t_i+l-2)$, причем $\|\xi_t\| < r$. Управления y_t в эти моменты избираются нулевыми. Далее траектория эволюционирует согласно уравнению. При нарушении ограничения $\|\xi_t\| \leq R$ или завершения текущего интервала Δ_i движение обрывается и вся процедура повторяется снова.

Обратимся к управлению процессами из класса \tilde{L} . Пусть сначала коэффициенты уравнения (2) известны. Предположим, что матрица B_1 невырожденная. Тогда оптимальное управление находится из условия

$$B_1 \boldsymbol{y}_t + B_2 \boldsymbol{y}_{t-1} + \dots + B_l \boldsymbol{y}_{t-l+1} + A_1 \xi_t + \dots + A_l \xi_{t-l+1} = 0,$$

т. е.

$$\begin{aligned} \boldsymbol{y}_t = -B_1^{-1} A_1 \xi_t - \dots - B_1^{-1} A_l \xi_{t-l+1} - \\ - B_2^{-1} B_1 \boldsymbol{y}_{t-1} - \dots - B_1^{-1} B_l \boldsymbol{y}_{t-l+1}. \end{aligned} \quad (4)$$

Если ввести $(n \times 2ln)$ -матрицу

$$M = (A_1, \dots, A_l, B_1, \dots, B_l)$$

и вектор $v_t = (\xi_t, \dots, \xi_{t-l+1}, y_t, \dots, y_{t-l+1})$, эту формулу можно записать короче:

$$B_1^{-1}Mv_t = 0.$$

Следовательно, $\xi_{t+1} = \zeta(t)$, и цель достигнута:

$$\|\xi_{t+1}\| = \|\zeta(t)\| < r.$$

Ограничения на управление здесь отсутствуют.

После этого предварительного замечания, делающего ясным структуру оптимального управления, обратимся к адаптивной системе для процессов класса L . Ясно, что в ней не может прямо использоваться формула (4), в которую входят неизвестные матрицы $B_1, \dots, B_l, A_1, \dots, A_l$.

Введем $(n \times 2ln)$ -матрицы

$$M_t = (A_1(t), \dots, A_l(t), B_1(t), \dots, B_l(t)),$$

считая, что они получены с помощью КСА и служат оценками матрицы M . Определим вектор

$$g_{t+1} = \xi_{t+1} - M_t v_t,$$

который, как и v_t , можно наблюдать в момент $t+1$ (после приложения управления y_t).

Допустим, что найдена такая матрица M_t , что выполнено неравенство

$$\|g_{t+1}\| = \|\xi_{t+1} - M_t v_t\| < r$$

и управление y_t выбрано из условия *)

$$M_t v_t = 0.$$

Ясно, что это обеспечивает достижение цели $\|\xi_{t+1}\| < r$. Однако следует еще учесть ограничения на управления $\|y_t\| \leq Q$.

Поставленную цель управления обеспечивает адаптивная система, в состав которой входят два блока — блок управления и блок оценивания. Определим их.

*) Оно подразумевает невырожденность матрицы $B_1(t)$ и отсутствие ограничений на управление.

В моменты действия блока управления («рабочие моменты») используется найденная на предыдущих этапах матрица M_t («оценка» матрицы M). С ее помощью очередные управление вычисляются по формулам, имеющим различный вид в зависимости от того, вырождена или нет матрица $B_1(t)$.

Если $B_1(t)$ невырождена,

$$\mathbf{y}_t = \begin{cases} U_t(M_t) & \text{при } \|U_t(M_t)\| < Q, \\ Q \frac{U_t(M_t)}{\|U_t(M_t)\|} & \text{при } \|U_t(M_t)\| \geq Q, \end{cases} \quad (5a)$$

где использовано обозначение $U_t(M_t) = -B_1^{-1}(t) M_t \mathbf{v}_t + \mathbf{y}_t$. Если же $B_1(t)$ вырождена, то найдем вектор \mathbf{e}_t единичной длины ($\|\mathbf{e}_t\| = 1$) такой, что $B_1(t) \mathbf{e}_t = 0$, и примем

$$\mathbf{y}_t = Q \mathbf{e}_t. \quad (5b)$$

Несложно убедиться, что эта формула является предельным случаем предыдущей, когда $\det B_1^{(t)}$ стремится к нулю.

Блок оценивания включается в моменты времени, следующие за рабочими и именуемые «моментами обучения», в это время управления полагаются равными нулю. Перед первым его включением задают какое-нибудь начальное значение — матрицу M_0 , а затем действует рекуррентная процедура

$$M_{t+1} = \begin{cases} M_t & \text{при } \|\mathbf{g}_{t+1}\| < r; \\ M_t + \frac{\mathbf{g}_{t+1} \mathbf{v}_t^T}{\|\mathbf{v}_t\|^2} & \text{при } \|\mathbf{g}_{t+1}\| \geq r; \end{cases} \quad (6)$$

\mathbf{v}_t^T означает транспонирование вектора-строки \mathbf{v}_t . Во все моменты времени, которые не есть моменты обучения, оценка не меняется. Отметим, что избранная здесь процедура недостаточна для достижения цели на всем классе процессов \tilde{L} . Впредь мы ограничиваемся подклассом $L_p \subset \tilde{L}$ таким, что

$$\|\zeta(t)\| < \rho r,$$

где $0 \leq \rho < 1/2$.

Займемся свойствами блока оценивания.

Теорема 2. Процедура (6) решения системы неравенств

$$\|\xi_{t+1} - M\mathbf{v}_t\| < r,$$

в которой управления строятся по (5а), (5б), является в классе L_p конечно-сходящимся алгоритмом, т. е. существует такой момент t^0 , что при всех $t \geq t^0$ неравенство выполнено и $M_t \equiv M_{t^0}$. Число исправлений алгоритма оценивается сверху величиной $l \frac{R^2 + Q^2}{(1 - 2\rho)r^2} \|M - M_0\|^2$.

Доказательство. Требуется установить убывание $\|M_t - M\|^2$. Введем числовую последовательность $\gamma_t = \|M_t - M\|^2 - \|M_{t+1} - M\|^2$. На каждом шаге изменения M_t (вторая формула в (6)) имеем

$$\gamma_t \geq \frac{(1 - 2\rho)r^2}{l(R^2 + Q^2)}.$$

Докажем это неравенство, опираясь на тождество $\|N + \mathbf{h}_1 \mathbf{h}_2^T\|^2 = \|N\|^2 + \|\mathbf{h}_1\|^2 \|\mathbf{h}_2\|^2$, где N — матрица порядка n , а \mathbf{h}_1 и \mathbf{h}_2 — векторы, причем $N\mathbf{h} = 0$. Будем писать $\mathbf{g}_{t+1}(B) = \xi_{t+1} - \Gamma\mathbf{v}_t$, где Γ — матрица такого же вида, как и M .

Обозначим $M' = M_{t+1} + \frac{[\mathbf{g}_{t+1}(M_t) - \mathbf{g}_{t+1}(M)] \mathbf{v}_t^T}{\|\mathbf{v}_t\|^2}$. В силу легко проверяемого равенства $\mathbf{g}_{t+1}(M') = \mathbf{g}_{t+1}(M)$ имеем $(M' - M)\mathbf{v}_t = 0$. Введем

$$\mathbf{N} = M' - M, \quad \mathbf{h} = \frac{\mathbf{v}_t}{\|\mathbf{v}_t\|^2},$$

тогда $\mathbf{N}\mathbf{h} = 0$. Воспользуемся упомянутым выше тождеством

$$\begin{aligned} \|M_t - M\|^2 &= \|N + (\mathbf{g}_{t+1}(M) - \mathbf{g}_{t+1}(M_t)) \mathbf{h}^T\|^2 = \\ &= \|N\|^2 + \|\mathbf{g}_{t+1}(M) - \mathbf{g}_{t+1}(M_t)\|^2 \|\mathbf{h}\|^2, \end{aligned}$$

кроме того,

$$\|M_{t+1} - M\|^2 = \|N + \mathbf{g}_{t+1}(M) \mathbf{h}^T\|^2 = \|N\|^2 + \|\mathbf{g}_{t+1}(M)\|^2 \|\mathbf{h}\|^2.$$

Отсюда находим

$$\begin{aligned} \gamma_t &= (\|\mathbf{g}_{t+1}(M) - \mathbf{g}(M_t)\|^2 - \|\mathbf{g}_{t+1}(M)\|^2) \|\mathbf{h}\|^2 = \\ &= \frac{\|\mathbf{g}_{t+1}(M_t)\|^2 - 2\mathbf{g}_{t+1}^T(M_t) \mathbf{g}_{t+1}(M)}{\|\mathbf{v}_t\|^2}. \end{aligned}$$

Из высказанных ранее предположений следуют неравенства

$$\|\mathbf{v}_t\|^2 = \sum_{i=1}^{l-1} (\|\xi_{t-i}\|^2 + \|\mathbf{y}_{t-i}\|^2) \leq l(R^2 + Q^2),$$

$$\|\mathbf{g}_{t+1}(M_t)\| \geq r, \quad \|\mathbf{g}_{t+1}(M)\| \leq \rho r.$$

Поэтому (с учетом свойства нормы произведения)

$$\begin{aligned} \gamma_t &\geq \|\mathbf{g}_{t+1}(M_t)\| (\|\mathbf{g}_{t+1}(M_t)\| - 2\|\mathbf{g}_{t+1}(M)\|) \|\mathbf{v}_t\|^{-2} \geq \\ &\geq r(r - \rho r)(l[R^2 + Q^2])^{-1}. \end{aligned}$$

Таким образом, установлено, что при каждом изменении M_t (по второй формуле в (6)) квадрат расстояния $\|M_t - M\|^2$ убывает не менее чем на $\delta = \frac{r^2(1-2\rho)}{l(R^2+Q^2)}$. Поэтому от M_0 до M можно сделать не свыше $\|M_0 - M\|^2/\delta$ шагов. Теорема доказана.

Остается показать, что блок управления непременно приведет процесс к назначенней цели. Перед формулировкой точного утверждения введем еще некоторые обозначения.

Предполагается, что число $x = \min_{z: \|z\|=1} \|B_1 z\| > 0$, c_i — положительные постоянные;

$$c_1 = \|M\|\sqrt{l} + 1 + \rho,$$

$$c_2 = \|M - M_0\|\sqrt{l + \nu^2 c_1^2(l-1)} + 1 + \rho,$$

$$c_3 = (\|M - M_0\| + \|M\|)\sqrt{l + \nu^2 c_1^2(l-1)}, \quad c_4 = \frac{c_2 + c_3}{x},$$

$$q = \epsilon \max(c_1, c_4).$$

Рассмотрим класс процессов $\tilde{L}_\rho \subset L_\rho$, для всех процессов из которого $x > 0$, $Q > q$.

Теорема 3. *Определенная выше система управления при всех $t \geq t^*$ приводит к достижению цели $\|\xi_t\| < r$. Для числа исправлений алгоритма сохраняется оценка из теоремы 2.*

Доказательство. По высказанному прежде условию все интервалы Δ_i конечные. Пусть момент t^0 , существование которого доказано в теореме 2, наступил на интервале Δ_k . Мы докажем, что моментом t^* служит

первый рабочий момент $t^{(k)}$ на интервале Δ_{k+1} . В действительности во все моменты этого интервала

$$\|\xi_t\| < r, \quad \|y_t\| < Q, \quad \|y_t\| \leq r c_1.$$

Рассуждения проведем лишь для момента $t^{(k)}$. Рассмотрим альтернативу: в этот момент ограничение $\|y_t\| < Q$ либо выполнено, либо нет. В первом случае целевое условие выполнено, так как сделанное предположение означает, что управление изображалось первым равенством в (5а). Это означает, что $B_1(t)$ невырождена и $M_t v_t = 0$. Отсюда и из того, что $\|g_{t+1}\| < r$ при $t \geq t^0$, следует выполнение цели. Далее в рассматриваемый (рабочий!) момент времени основное уравнение (2) можно записать следующим образом:

$$B_1 y_t + \dots + B_l y_{t-l+1} = z_t,$$

где $z_t = \xi_{t+1} - A_1 \xi_t - \dots - A_l \xi_{t-l+1} - \zeta(t)$.

Из последнего равенства имеем

$$\|z_t\| < r + \|M\| \sqrt{l} r + r \rho = r c_1$$

и остается воспользоваться свойством диссипативности. Для этого следует положить в лемме $m=0$, $c=r c_1$, $d=r c_1$.

Установим теперь, что в момент $t^{(k)}$ не может выполниться равенство $\|y_t\|=Q$. Доказываем способом от противного, а это значит, что управление вычислялось либо а) по второму равенству (5а), если $\det B_1(t) \neq 0$, либо б) по (5б), если $\det B_1(t)=0$. Неравенство $\|g_{t+1}(M_t)\| < r$ запишем так: $\|(B_1 - B_1(t)) y_t + h_t(M - M_t) + \zeta(t)\| < r$, где принято для краткости обозначение $h_t(M - M_t) = (M - M_t) v_t - (B_1 - B_1(t)) y_t$. Оценим норму последней величины

$$\begin{aligned} \|h_t\|^2 &\leq \left[\sum_{i=1}^l \|A_i - A_i(t)\|^2 + \sum_{i=2}^l \|B_i - B_i(t)\|^2 \right] \times \\ &\quad \times \left[\sum_{i=1}^l \|\xi_{t-i}\|^2 + \sum_{i=2}^l \|y_{t-i}\|^2 \right], \end{aligned}$$

но согласно ранее сказанному

$$\|\xi_{t-i}\| < r, \quad i = \overline{1, l}, \quad \|y_{t-i}\| = 0 \leq r c_1, \quad i = \overline{2, l},$$

$$\|M - M_t\| \leq \|M - M_0\|.$$

Поэтому получаем

$$\|\mathbf{h}_t\| \leq r \|M - M_0\| \sqrt{l + v^2 c_1^2(l-1)},$$

а следовательно,

$$\|(B_1 - B_1(t)) \mathbf{y}_t\| \leq r c_2. \quad (7)$$

Для оценки нормы $B_1(t) \mathbf{y}_t$ в варианте управления а) имеем $B_1(t) \mathbf{y}_t = a(t) [M_t \mathbf{v}_t - B_1(t) \mathbf{y}_t]$, где $a(t) = \frac{Q}{\|U_t(M_t)\|} \leq 1$, а в варианте б) $B_1(t) \mathbf{y}_t = 0$. Аналогично предыдущему находим

$$\|\mathbf{h}_t(A_t)\| \leq r (\|M_0 - M\| + \|M\|) \sqrt{l + v^2 c_1^2(l-1)} = r c_3.$$

Таким образом, в обоих вариантах $\|B_1(t) \mathbf{y}_t\| \leq r c_3$. Если теперь в (7) умножить выражение под знаком нормы слева на B_1^{-1} , то получим $\|\mathbf{y}_t - B_1^{-1} B_1(t) \mathbf{y}_t\| \leq \frac{r c_2}{x}$.

а отсюда, наконец, получаем $\|\mathbf{y}_t\| \leq r \frac{1}{x} (c_2 + c_3) = r c_4$.

По сделанному предположению $Q > r c_4$, значит, приходим к $\|\mathbf{y}_t\| < Q$, а не к $\|\mathbf{y}_t\| = Q$. Следовательно, приведенное только что неравенство не может нарушиться в момент $t^{(k)}$. Так же рассуждаем и для всех последующих моментов $(t^{(k)} + j, j \geq 1)$. Рассмотрения для них отличаются от изложенных тем, что теперь $y_{t-i} \neq 0, i = 2, \dots, l$, при $t > t^{(k)}$. Однако обратим внимание на то, что в предыдущем доказательстве использовались неравенства $\|\mathbf{y}_{t-i}\| \leq r v c_1$, а они справедливы всегда. Рассуждения сохраняют силу и для дальнейших интервалов $\Delta_{k+2}, \Delta_{k+3}, \dots$ Теорема доказана.

Заметим, что оценка момента времени, начиная с которого целевое неравенство будет выполнено всегда, не может быть дана в общем случае, ибо момент последнего «возмущения» процесса возможен в сколь угодно далеком будущем. Оценки фактически встречающихся количеств исправлений алгоритма возможно получить посредством имитации на ЦВМ объекта и адаптивной системы. Проведенные расчеты показали, что обычно количества исправлений умеренные.

ГЛАВА IX

УПРАВЛЕНИЕ ПРОЦЕССАМИ ОБЩЕГО ВИДА

§ 1. Стационарные процессы

Предметом исследований в этой главе являются адаптивные системы для управления процессами общего вида, т. е. такими, у которых условные распределения $\mu_{t+1}(\cdot | x^t, y^t)$ явным образом зависят от всей предшествующей эволюции. Естественно, что на систему распределений μ_t , необходимо накладывать достаточно жесткие ограничения, которые обеспечивали бы существование желаемых адаптивных систем. Так, в относительно простом случае марковских процессов приходится требовать эргодичность.

В настоящем параграфе рассматриваются управляемые случайные процессы, определенные на множестве всех целых чисел, т. е. подразумеваем, что время t пробегает все отрицательные и положительные значения $t = \dots, -2, -1, 0, 1, 2, \dots$. В соответствии с этим примем обозначение

$$z^t = (\dots, z_{t-2}, z_{t-1}, z_t)$$

для бесконечной влево последовательности, в частности

$$z^0 = (\dots, z_{-2}, z_{-1}, z_0).$$

Как и прежде, X и Y означают измеримые пространства фазовое и управлений. Управляемые условные распределения записываются в виде $\mu_{t+1}(M | x^t, y^t)$.

Управляемым стационарным процессом называется управляемый случайный процесс ξ_t , условные распределения которого при любом t подчинены условию

$$\mu(M | x^t, y^t) \equiv \mu(M | x^0, y^0),$$

если $x^t = x^0$ и $y^t = y^0$.

Высказанное условие означает, что вероятности $\mu(M|x^t, y^i)$, $M \in \mathfrak{M}$, инвариантны относительно сдвига времени.

Процессы подобного типа возникают в задачах передачи и обработки информации, управления системами связи. Общеизвестна роль стационарных (неуправляемых) процессов в теории регулирования, в которой возникают проблемы прогнозирования течения таких процессов, фильтрации полезных сигналов от помех и т. п. Следует обратить внимание, что свойство стационарности требует рассмотрения процессов на бесконечном в обе стороны временном множестве.

Конкретные случайные процессы будем получать из управляемого стационарного процесса посредством стационарных стратегий произвольной конечной глубины. Такие стратегии задаются одной измеримой функцией $\sigma_m: X^m \rightarrow Y$. Значение функции в момент t определено равенством $y_t = \sigma_m(\xi_{t-m}^{t-1})$, из которого видно, что правило выбора очередного действия не зависит от предыдущих действий, а сама функция явным образом не зависит от времени. Множество стационарных стратегий глубины m обозначим D_m и положим

$$D = \bigcup_m D_m.$$

Пусть f — произвольная ограниченная измеримая функция на X . Введем обозначение для величины общего предела, если он существует *),

$$\lim_{t \rightarrow \infty} \mathbf{E} f(\xi_t) = \text{l. i. m.}_{T \rightarrow \infty} \frac{1}{T} \sum_{i=1}^T f(\xi_i) = W(\sigma).$$

Кроме того, положим

$$\bar{W} = \sup_{\sigma} W(\sigma),$$

*) Символ л. и. м. означает предел в среднем (квадратическом), т. е.

$$\lim_{T \rightarrow \infty} \mathbf{E} \left(\frac{1}{T} \sum_{i=1}^T f(\xi_i) - W(\sigma) \right)^2 = 0.$$

где верхняя грань берется по множеству допустимых стратегий D .

Целями управления классом управляемых стационарных процессов назначим соответственно достижение равенств:

$$1) \quad \text{l. i. m. } \frac{1}{T} \sum_1^T f(\xi_t) = \bar{W},$$

$$2) \quad \lim_{t \rightarrow \infty} E f(\xi_t) = \bar{W},$$

причем фигурирующие здесь пределы должны существовать. Заранее можно ожидать, что стратегии, которые обеспечивают сформулированные цели на классе процессов, нестационарны.

Сделаем теперь основное (для способа изложения) допущение, что пространствами X и Y служат конечные отрезки на числовой прямой.

Лебеговскую меру пространства X обозначим через $\text{mes}(X) = \mu$.

Предположим, что условные распределения процесса имеют плотности вероятностей

$$p(x|x^t, y^t).$$

Пусть для процесса выполнено условие: при всех x, x^t, y

$$0 < \delta_1 \leq p(x|x^t, y^t) \leq \delta_2 < \infty.$$

Синтезу адаптивных систем для рассматриваемых классов процессов предпошим исследование стационарных процессов (без управления), задаваемых инвариантными относительно сдвигов по t условными плотностями вероятностей $p(x|x^t)$. Пусть они удовлетворяют тем же неравенствам, что и $p(x|x^t, y^t)$.

Нас будут интересовать следующие характеристики этих процессов:

$$\varepsilon_k = \sup_{x_{-k}^1, x_{-k}^{-k-1}, \hat{x}_{-k-k-1}} |p(x_1|x_{-k}^0, x^{-k-1}) - p(x_1|x_{-k}^0, \hat{x}^{-k-1})|,$$

$$A_k = \sup_{x_t, t, x_{-k}^0, x_{-k}^{-k-1}, \hat{x}_{-k-k-1}} |p(x_t|x_{-k}^0, x^{-k-1}) - p(x_t|x_{-k}^0, \hat{x}^{-k-1})|.$$

Числа ε_k и A_k оценивают зависимость вероятностных свойств процесса от его предыстории за последние $k+1$ тактов.

Лемма 1. Если $\sum_{k=1}^{\infty} \varepsilon_k < \infty$, то $\lim_{k \rightarrow \infty} A_k = 0$.

При доказательстве используем вспомогательные величины

$$a_{t,k} = \sup_{x_t, x^0, \tilde{x}^{-k-1}} | p(x_t | x_{-k}^0, x^{-k-1}) - p(x_t | x_{-k}^0, \tilde{x}^{-k-1}) |,$$

связанные с A_k равенством $A_k = \sup_t a_{t,k}$, а также неравенство

$$\begin{aligned} p(x_t | x^1) &= \int_{X^{t-2}} p(x_t | x^{t-1}) p(x_{t-1} | x^{t-2}) \dots \\ &\quad \dots p(x_2 | x^1) dx_{t-1} \dots dx_2 \leqslant \\ &\leqslant \delta_2 \int_{X^{t-2}} p(x_{t-1} | x^{t-2}) \dots p(x_2 | x^1) dx_{t-1} \dots dx_2 = \delta_2. \end{aligned}$$

Имеем

$$\begin{aligned} a_{t,k} &= \sup_{x_t, x^0, \tilde{x}^{-k-1}} \left| \int_X p(x_t | x_{-k}^1, x^{-k-1}) p(x_1 | x_{-k}^0, x^{-k-1}) dx_1 - \right. \\ &\quad \left. - \int_X p(x_t | x_{-k}^1, \tilde{x}^{-k-1}) p(x_1 | x_{-k}^0, \tilde{x}^{-k-1}) dx_1 \right| \leqslant \\ &\leqslant \sup \left| \int_X p(x_t | x_{-k}^1, x^{-k-1}) | p(x_1 | x^0) - \right. \\ &\quad \left. - p(x_1 | x_{-k}^0, \tilde{x}^{-k-1}) | dx_1 + \int_X p(x_1 | x_{-k}^0, \tilde{x}^{-k-1}) | p(x_t | x^1) - \right. \\ &\quad \left. - p(x_t | x_{-k}^1, \tilde{x}^{-k-1}) | dx_1 \right| \leqslant \sup_X [p(x_t | x^1) \varepsilon_k + \right. \\ &\quad \left. + p(x_1 | x_{-k}^0, \tilde{x}^{-k-1}) a_{t-1, k+1}] dx_1 \leqslant \delta_2 \mu \varepsilon_k + a_{t-1, k+1}. \end{aligned}$$

Интегрируя это неравенство, приходим к следующему:

$$a_{t,k} = \delta_2 \mu (\varepsilon_k + \varepsilon_{k+1} + \dots + \varepsilon_{k+t-2}) + a_{1, k+t-1} =$$

$$= \delta_2 \mu (\varepsilon_k + \dots + \varepsilon_{k+t-2}) + \varepsilon_{k+t-1} \leqslant \delta_2 \mu \sum_{i=k}^{k+t-1} \varepsilon_i.$$

Значит, $A_k = \sup_t a_{t,k} \leq \delta_2 \mu \sum_{i=k}^{\infty} \varepsilon_i$, и в случае сходимости ряда $\sum_{i=1}^{\infty} \varepsilon_i$ имеем $\lim_{k \rightarrow \infty} A_k = 0$.

Лемма 2. Если

$$\lim_{k \rightarrow \infty} \frac{\varepsilon_k}{(\delta_1 \mu)^k} = 0,$$

то при любой последовательности x^0 существует предельная плотность вероятностей $p(x)$, не зависящая от x^0 ,

$$\lim_{t \rightarrow \infty} p(x_t = x | x^0) = p(x).$$

Доказательство проводится оценкой разности $M_t(x) - m_t(x)$, где $M_t(x) = \sup_{x^0} p(x_t = x | x^0)$ — невозрастающая, а $m_t(x) = \inf_{x^0} p(x_t = x | x^0)$ — неубывающая функции аргумента t .

Существует такая монотонно стремящаяся к 0 функция $\varphi(t)$, что

$$|p(x_t = x | x^0) - p(x)| \leq \varphi(t).$$

Лемма 3. Если $\lim_{k \rightarrow \infty} \frac{\varepsilon_k}{(\delta_1 \mu)^k} = 0$, то существует не зависящий от x^0 предел

$$\lim_{T \rightarrow \infty} E(f(x_T) | x^0) = \int_X f(x) p(x) dx,$$

причем

$$\left| \int_X f(x) p(x) dx - E(f(x_T) | x^0) \right| \leq \varphi(T) \int_X |f(x)| dx.$$

Лемма 4. При условии леммы 3

$$\text{l. i. m}_{T \rightarrow \infty} \frac{1}{T} \sum_{i=1}^T f(x_i) = \int_T f(x) p(x) dx,$$

причем

$$\mathbb{E} \left| \frac{1}{T} \sum_{i=1}^T f(x_i) - \int_X f(x) p(x) dx \right|^2 \leq \psi(T) = \frac{4F^2}{T} \left(1 + \mu \sum_{i=1}^T \varphi(i) \right),$$

где $F = \sup_x |f(x)|$, $\lim_{T \rightarrow \infty} \psi(T) = 0$.

Эти результаты мы перенесем теперь на управляемые стационарные последовательности. Детали доказательств, а в простых случаях и схемы доказательств опускаются.

Введем обозначение $\xi_t(\sigma)$ для управляемого стационарного процесса, находящегося под воздействием стратегии $\sigma \in D$.

Лемма 5. *При всякой стратегии $\sigma \in D$ и любом стационарном процессе ξ_t процесс $\xi_t(\sigma)$ является стационарным случайным процессом.*

Введем, подобно тому как это было сделано для случайных процессов, параметры

$$\varepsilon_k = \sup |p(x_1 | x^0, y^0) - p(x_1 | x_{-k}^0, \tilde{x}^{-k-1}, y_{-k}^0, \tilde{y}^{-k-1})|,$$

$$A_k = \sup |p(x_t | x^0, y^0) - p(x_t | x_{-k}^0, \tilde{x}^{-k-1}, y_{-k}^0, \tilde{y}^{-k-1})|$$

(верхние грани берутся по всем входящим в выражения переменным), а для процесса $\xi_t(\sigma)$, $\sigma \in D_m$ — параметры $\varepsilon_{k,m}$, $A_{k,m}$, а также нижние и верхние оценки плотности $\delta_{1,m}$, $\delta_{2,m}$.

Лемма 6. *Для любого целого k и глубины m стратегии σ*

$$\varepsilon_{k,m} \leq \varepsilon_{k-m}, \quad \delta_{1,m} \geq \delta_1, \quad \delta_{2,m} \leq \delta_2.$$

Лемма 7. *Если для управляемого процесса ξ_t выполнено равенство $\lim_{k \rightarrow \infty} \frac{\varepsilon_k}{(\delta_{1,m})^k} = 0$, то для процесса $\xi_t(\sigma)$, $\sigma \in D_m$, имеет место*

$$\lim_{k \rightarrow \infty} \frac{\varepsilon_{k,m}}{(\delta_{1,m})^k} = 0.$$

Из этого результата и лемм 2—4 можно вывести следующие три утверждения:

Лемма 8. Для каждого целого t существует такая монотонная положительная функция $\varphi_m(t)$, $\lim_{t \rightarrow \infty} \varphi_m(t) = 0$, что при $\sigma \in D_m$

$$|p_\sigma(x_t = x | x^0, y^0) - p_\sigma(x)| \leq \varphi_m(t).$$

Лемма 9. Если $\lim_{k \rightarrow \infty} \frac{\epsilon_k}{(\delta_1 \mu)^k} = 0$, то при любой стратегии $\sigma \in D_m$ существует не зависящий от (x^0, y^0) предел

$$\lim_{T \rightarrow \infty} E_\sigma(f(x_T) | x^0, y^0) = \int_X f(x) p_\sigma(x) dx = W(\sigma),$$

причем

$$\left| \int_X f(x) p(x) dx - E(f(x_T) | x^0, y^0) \right| \leq \varphi_m(T) \int_X |f| dx.$$

Лемма 10. Если $\lim_{k \rightarrow \infty} \frac{\epsilon_k}{(\delta_1 \mu)^k} = 0$, то для всякой стратегии $\sigma \in D_m$ существует не зависящий от (x^0, y^0) предел

$$\text{l. i. m. } \frac{1}{T} \sum_{t=1}^T f(\xi_t(\sigma)) = \int_X f(x) p_\sigma(x) dx,$$

причем

$$\begin{aligned} E\left(\left[\frac{1}{T} \sum_{t=1}^T f(\xi_t(\sigma)) - \int_X f(x) p_\sigma(x) dx\right]^2 \middle| x^0, y^0\right) &\leq \psi_m(T) = \\ &= \frac{4F^2}{T} \left(1 + \mu \sum_{i=1}^T \varphi_m(i)\right), \end{aligned}$$

где

$$F = \sup_x |f(x)|, \quad \lim_{T \rightarrow \infty} \psi_m(T) = 0.$$

В множестве D_m выделим какое-нибудь конечное подмножество D'_m и для каждой стратегии $\sigma \in D'_m$ определим случайную величину $z_\sigma = \frac{1}{n} \sum_{t=1}^n f(\xi_t(\sigma))$. Пусть σ_m^0 — наилуч-

шая во множество D'_m стратегия, т. е.

$$\max_{\sigma \in D'_m} W(\sigma) = W(\sigma_m^0).$$

Через $\tilde{\sigma}$ обозначим ту стратегию (случайную!), которой отвечает максимальная величина z_σ . О «расстоянии» между этими стратегиями говорит следующая лемма.

Лемма 11. *При любой предыстории (x^0, y^0) выполняется неравенство*

$$E([W(\tilde{\sigma}) - W(\sigma_m^0)]^2 | x^0, y^0) \leq 4 |D'_m| \psi_m(n).$$

Используем еще такие обозначения: пусть $W_m = \sup_{\sigma \in D_m} W(\sigma)$, эти числа образуют неубывающую последовательность. Поэтому величина $W^0 = \sup_{\sigma \in D} W(\sigma)$ является пределом последовательности

$$W^0 = \lim_{m \rightarrow \infty} W_m.$$

Определим вспомогательные множества стратегий D_m^l , $l = 1, 2, \dots$. Для этого разобьем множества X и Y на l равных интервалов. Их длины равны соответственно $\Delta x_l = |X|/l$ и $\Delta y_l = |Y|/l$, а точки деления x^0, x^1, \dots, x^l и y^0, y^1, \dots, y^l (x^0, y^0 — начальные, x^l, y^l — концевые). Рассмотрим функции $f(x_1, \dots, x_m)$, кусочно постоянные в гиперкубе X^m : в его каждой области с границами $(x_1^{i_1}, x_1^{i_1+1}), (x_2^{i_2}, x_2^{i_2+1}), \dots, (x_m^{i_m}, x_m^{i_m+1})$ функция принимает одно из значений y^i , $i = 0, 1, \dots, l$. Множество D_m^l образовано из всех таких функций.

Как известно, произвольную измеримую функцию можно аппроксимировать кусочно постоянными функциями. В частности, задав числа $\eta_1 > 0$ и $\eta_2 > 0$, можно выделить подмножество $X_{\eta_2} \subset X$ такое, что $\text{mes } X_{\eta_2} < \eta_2$ и при всех $x \in X \setminus X_{\eta_2}$ имеем $|\sigma - \sigma'| < \eta_1$, где $\sigma \in D_m$, а $\sigma' \in D_m^l$. На этом факте базируется дальнейшее изложение.

Пусть $p(x_1 | x_{-k+1}^0)$ — плотность переходной функции k -связного марковского процесса с фазовым пространством X , удовлетворяющая неравенствам

$$0 < \delta_1 \leq p(x_1 | x_{-k+1}^0) \leq \delta_2 < \infty. \quad (1)$$

Кроме того, зададим плотности $p'(x_1 | x_{-k+1}^0)$ и $\tilde{p}(x_1 | x_{-k+1}^0)$, подчиненные этому же условию, а также

$$|p(x_1 | x_{-k+1}^0) - p'(x_1 | x_{-k+1}^0)| < \eta_1$$

для всех $x_1, x_0, x_{-1}, \dots, x_{-k+1} \in X$ и

$$p'(x_1 | x_{-k+1}^0) = \tilde{p}(x_1 | x_{-k+1}^0)$$

для всех $x_i \in X, x_i \in X \setminus X_{\eta_2}, i = 0, -1, \dots, -k+1$, где $\text{mes } X_{\eta_2} = \eta_2$. Все три перечисленных процесса имеют плотности предельных распределений $p(x)$, $p'(x)$ и $\tilde{p}(x)$ соответственно. Ближайшая наша цель — оценить $\|p - \tilde{p}\|_{L_1}$.

Лемма 12.

$$\|p - p'\|_{L_1} \leq \eta_1 k^{\frac{\delta_2^{k-1}}{\delta_1^k}}.$$

Доказательство. Интересующие нас предельные распределения рассмотрим на наборах x_1^k . Имеем

$$p(x_{k+1}^{2k}) = \int_{X^k} p(x_{k+1}^{2k} | x_1^k) p(x_1^k) dx_1^k = Tp,$$

$$p'(x_{k+1}^{2k}) = \int_{X^k} p'(x_{k+1}^{2k} | x_1^k) p'(x_1^k) dx_1^k = T'p'.$$

Значит, для чисел $I = \|p(x_{k+1}^{2k}) - p'(x_{k+1}^{2k})\|_{L_1^{(k)}}$, $L_1^k = L_1(R^k)$ имеем

$$I = \|Tp - T'p\|_{L_1^{(k)}} \leq \|Tp - T'p\|_{L_1^{(k)}} + \|T'p - T'p'\|_{L_1^{(k)}}.$$

Сначала оценим первое слагаемое справа

$$\|Tp - T'p\|_{L_1^{(k)}} \leq \int_{X^k} \int_{X^k} |p(x_{k+1}^{2k} | x_1^k) - p'(x_{k+1}^{2k} | x_1^k)| p(x_1^k) dx_1^k dx_{k+1}^{2k}.$$

Рассмотрим модуль под знаком интеграла. В силу равенства

$$p(x_{k+1}^{2k} | x_1^k) = p(x_{2k} | x_1^{2k-1}) p(x_{2k-1} | x_1^{2k-2}) \dots p(x_{k+1} | x_1^k)$$

и аналогичного для p' , а также элементарного неравенства ($a_i \geq 0, b_i \geq 0$)

$$\begin{aligned} |a_1 a_2 \dots a_k - b_1 b_2 \dots b_k| &\leq |a_1 - b_1| a_2 \dots a_k + \\ &+ b_1 |a_2 - b_2| a_3 \dots a_k + \dots + b_1 b_2 \dots b_{k-1} |a_k - b_k|, \end{aligned}$$

находим

$$|p(x_{k+1}^{2k} | x_1^k) - p'(x_{k+1}^{2k} | x_1^k)| \leq \eta_1 k \delta_2^{k-1}.$$

Отсюда следует, что

$$\|Tp - T'p\|_{L_1^{(k)}} \leq \eta_1 k \delta_2^{k-1} \int_{X^k} \int_{X^k} p(x_1^k) dx_1^k dx_{k+1}^{2k} = \eta_1 k \mu^k \delta_2^{k-1}.$$

Займемся теперь вторым слагаемым. Введем функцию $p_{\delta_1}(x_{k+1}^{2k} | x_1^k) = p(x_{k+1}^{2k} | x_1^k) - \delta_1^k \geq 0$. Имеем

$$\begin{aligned} \|T'p - T'p'\|_{L_1^{(k)}} &\leq \int_{X^k} \left| \int_{X^k} (p(x_1^k) - p'(x_1^k)) p'(x_{k+1}^{2k} | x_1^k) dx_1^k \right| dx_{k+1}^{2k} = \\ &= \int_{X^k} \left| \int_{X^k} p'_{\delta_1}(x_{k+1}^{2k} | x_1^k) (p(x_1^k) - p'(x_1^k)) dx_1^k \right| dx_{k+1}^{2k} \leq \\ &\leq \int_{X^k} \left[\int_{X^k} p'_{\delta_1}(x_{k+1}^{2k} | x_1^k) dx_{k+1}^{2k} \right] |p(x_1^k) - p'(x_1^k)| dx_1^k = \\ &= (1 - (\mu \delta_1)^k) \|p - p'\|_{L_1^{(k)}}. \end{aligned}$$

Из полученной оценки находим

$$I = \|p - p'\|_{L_1^{(k)}} \leq \eta_1 k \mu^k \delta_2^{k-1} + (1 - (\mu \delta_1)^k) \|p - p'\|_{L_1^{(k)}},$$

откуда

$$\|p - p'\|_{L_1^{(k)}} \leq \eta_1 k \frac{\delta_2^{k-1}}{\delta_1^k}.$$

Остается заметить, что одномерная предельная плотность равна $p(x_k) = \int_{X^{k-1}} p(x_1^k) dx_1^{k-1}$. Поэтому для норм справедливо неравенство

$$\begin{aligned} \|p - p'\|_{L_1} &= \int_X \left| \int_{X^{k-1}} [p(x_1^k) - p'(x_1^k)] dx_1^{k-1} \right| dx_k \leq \\ &\leq \int_{X^k} |p(x_1^k) - p'(x_1^k)| dx_1^k = \|p - p'\|_{L_1^{(k)}}, \end{aligned}$$

которое влечет справедливость утверждения леммы.

Лемма 13.

$$\| p' - \tilde{p} \|_{L_1} \leq (2k-1) \frac{\eta_2 \delta_2^{2k} \mu^{k-1}}{\delta_1^k}.$$

Доказательство. Снова рассматриваем плотности $p'(x_1^k)$ и $\tilde{p}(x_1^k)$ и равенства для них

$$p' = T' p', \quad \tilde{p} = \tilde{T} \tilde{p},$$

где T' и \tilde{T} — интегральные операторы с ядрами $p'(x_{k+1}^{2k} | x_1^k)$ и $\tilde{p}(x_{k+1}^{2k} | x_1^k)$ соответственно. Имеем

$$\begin{aligned} I &= \| p' - \tilde{p} \|_{L_1^{(k)}} = \\ &= \| T' p' - \tilde{T} \tilde{p} \|_{L_1^{(k)}} \leq \| T' p' - \tilde{T} p' \|_{L_1^{(k)}} + \| \tilde{T} p' - \tilde{T} \tilde{p} \|_{L_1^{(k)}}. \end{aligned}$$

Оценим сначала первое слагаемое справа

$$\begin{aligned} \| T' p' - \tilde{T} p' \|_{L_1^{(k)}} &\leq \\ &\leq \int \int | p'(x_{k+1}^{2k} | x_1^k) - \tilde{p}(x_{k+1}^{2k} | x_1^k) | p'(x_1^k) dx_1^k dx_{k+1}^{2k}. \end{aligned}$$

Равенство $p'(x_{k+1}^{2k} | x_1^k) = \tilde{p}(x_{k+1}^{2k} | x_1^k)$ выполняется, если одновременно при всех $i = 1, \dots, 2k-1$ оказывается $x_i \in X - X_{\eta_2}$. Мера множества всех наборов x_1^{2k} , для которых не выполняются эти включения, равна

$$\mu^{2k} \left[1 - \left(1 - \frac{\eta_2}{\mu} \right)^{2k-1} \right] \leq (2k-1) \eta_2 \mu^{2k-1}.$$

Значит,

$$\| T' p' - \tilde{T} p' \|_{L_1^{(k)}} \leq (2k-1) \eta_2 \mu^{2k-1} \delta_2^{2k}.$$

Второе слагаемое оценивается аналогично предыдущей лемме, т. е.

$$\| T' p' - \tilde{T} \tilde{p} \|_{L_1^{(k)}} \leq [1 - (\mu \delta_1)^k] \| p' - \tilde{p} \|_{L_1^{(k)}}.$$

Отсюда получаем

$$I = \| p' - \tilde{p} \|_{L_1^{(k)}} \leq (2k-1) \eta_2 \mu^{2k-1} \delta_2^{2k} + [1 - (\mu \delta_1)^k] \| p' - \tilde{p} \|_{L_1^{(k)}},$$

т. е. окончательно

$$\| p' - \tilde{p} \|_{L_1^{(k)}} \leq (2k-1) \frac{\eta_2 \delta_2^{2k} \mu^{k-1}}{\delta_1^k}.$$

Как и в лемме 12, приходим к желаемой оценке для $\|p' - \tilde{p}\|_{L_1}$.

На основании лемм 12 и 13 получается лемма 14.

Лемма 14.

$$\|p - \tilde{p}\|_{L_1} \leq \eta_1 k^{\frac{\delta_2^{k-1}}{\delta_1^k}} + (2k-1) \eta_2 \frac{\delta_2^{2k} \mu^{k-1}}{\delta_1^k}.$$

Снова рассмотрим плотность условного распределения $p(x_1 | x^0)$ стационарного случайного процесса и плотность переходной функции $p(x_1 | x_{-k}^0)$ k -связного марковского процесса, подчиненную условию

$$\inf_{x_{-k-1}} p(x_1 | x_{-k}^0, x^{-k-1}) \leq p'(x_1 | x_{-k}^0) \leq \sup_{x_{-k-1}} p(x_1 | x_{-k}^0, x^{-k-1}).$$

Лемма 15. При любой предыстории x^0 введенные плотности связаны неравенством

$$|p(x_t | x_{-k}^0) - p'(x_t | x^0)| \leq \varepsilon_{k-1} \delta_2 \mu t.$$

Лемма 16. Пусть $p(x)$ — плотность предельного распределения k -связного марковского процесса с плотностью $p(x_1 | x_{-k}^0)$ переходной функции. Тогда

$$|p(x_t = x | x_{-k}^0) - p(x)| \leq C [1 - (\mu \delta_1)^k]^{t/k},$$

где C — постоянная, не зависящая от k .

Подчиним плотности условных распределений управляемого случайного процесса ξ_t следующему условию:

Для любых двух последовательностей управлений y^0 и \tilde{y}^0

$$|p(x_1 | x^0, y^0) - p(x_1 | x^0, \tilde{y}^0)| \leq \sum_{i=1}^{\infty} d_i |y_{-i} - \tilde{y}_{-i}|, \quad (2)$$

где $d_i > 0$ и $d = \sum_{i=1}^{\infty} d_i < \infty$.

Теорема 1. Пусть $\sigma_m^{l, 0}$ — наилучшая стратегия из D_m^l . При соблюдении условий (2) и

$$\lim_{k \rightarrow \infty} k \varepsilon_k \left(\frac{\delta_2}{\delta_1} \right)^k = 0$$

для любой возрастающей целочисленной последовательности $l(m)$ выполняются соотношения

$$\lim_{l \rightarrow \infty} W(\sigma_m^{l,0}) = \lim_{l \rightarrow \infty} \max_{\sigma \in D_m^{l(m)}} W(\sigma) = W_m$$

($\max_{\sigma \in D_m^{l(m)}} W(\sigma)$ будем обозначать через W_m^l).

Доказательство. Пусть для всякой стратегии $\sigma \in D_m$ по заданным числам $\eta_1 > 0$ и $\eta_2 > 0$ найдены целое $l = l(m)$ и стратегия $\tilde{\sigma} \in D_m^l$ такая, что

$$|\sigma(x_{-1}, x_{-2}, \dots, x_{-m}) - \tilde{\sigma}(x_{-1}, x_{-2}, \dots, x_{-m})| \leq \eta_1$$

для всех $x_{-i} \in X - X_{\eta_2}$, i, \dots, m . Тогда

$$|p_\sigma(x_1 | x^0) - p_{\tilde{\sigma}}(x_1 | x^0)| \leq \sum_{i=1}^{\infty} d_i |y_{-i} - \tilde{y}_{-i}| \leq d\eta_1$$

для всех указанных только что x_{-i} .

Сопоставим процессам $\xi_t(\sigma)$ и $\xi_t(\tilde{\sigma})$ k -связные марковские процессы с плотностями $p_\sigma^{(k)}(x_1 | x_{-k+1}^0)$ и $p_{\tilde{\sigma}}^{(k)}(x_1 | x_{-k+1}^0)$, подчиненными условиям

$$\inf_{x_1, x_{-k+1}^0} p_\sigma(x_1 | x_{-k+1}^0) \leq p_\sigma^{(k)}(x_1 | x_{-k+1}^0) \leq \sup_{x_1, x_{-k+1}^0} p_\sigma(x_1 | x_{-k+1}^0),$$

$$\inf_{x_1, x_{-k+1}^0} p_{\tilde{\sigma}}(x_1 | x_{-k+1}^0) \leq p_{\tilde{\sigma}}^{(k)}(x_1 | x_{-k+1}^0) \leq \sup_{x_1, x_{-k+1}^0} p_{\tilde{\sigma}}(x_1 | x_{-k+1}^0).$$

Следующие неравенства очевидны:

$$\begin{aligned} |p_\sigma^{(k)}(x_1 | x_{-k+1}^0) - p_\sigma(x_1 | x^0)| &\leq \varepsilon_{k-m-1}, \\ |p_{\tilde{\sigma}}^{(k)}(x_1 | x_{-k+1}^0) - p_{\tilde{\sigma}}(x_1 | x^0)| &\leq \varepsilon_{k-m-1}. \end{aligned} \tag{3}$$

При высказанных предположениях существуют предельные плотности вероятности. Их мы будем отличать от «допредельных» плотностей аргументами x_∞ . Оценим следующую норму:

$$\begin{aligned} \|p_\sigma(x_\infty | x^0) - p_{\tilde{\sigma}}(x_\infty | x^0)\|_{L_1} &\leq \|p_\sigma(x_\infty | x^0) - \\ &- p_\sigma(x_t | x^0)\|_{L_1} + \|p_\sigma(x_t | x^0) - p_\sigma^{(k)}(x_t | x_{-k+1}^0)\|_{L_1} + \\ &+ \|p_\sigma^{(k)}(x_t | x_{-k+1}^0) - p_{\tilde{\sigma}}^{(k)}(x_\infty | x_{-k+1}^0)\|_{L_1} + \end{aligned}$$

$$\begin{aligned}
 & + \| p_{\sigma}^{(k)}(x_{\infty} | x_{-k+1}^0) - p_{\delta}^{(k)}(x_{\infty} | x_{-k+1}^0) \|_{L_1} + \\
 & + \| p_{\delta}^{(k)}(x_{\infty} | x_{-k+1}^0) - p_{\delta}^{(k)}(x_t | x_{-k+1}^0) \|_{L_1} + \\
 & + \| p_{\delta}^{(k)}(x_t | x_{-k+1}^0) - p_{\delta}(x_t | x^0) \|_{L_1} + \\
 & + \| p_{\delta}(x_t | x^0) - p_{\delta}(x_{\infty} | x^0) \|_{L_1}. \quad (4)
 \end{aligned}$$

Оценим поодиночке слагаемые в правой части этого неравенства.

Из соотношений (2) и (3) следует, что для $x_1 \in X$

$$\begin{aligned}
 |p_{\sigma}^{(k)}(x_1 | x_{-k+1}^0) - p_{\delta}^{(k)}(x_1 | x_{-k+1}^0)| & \leq |p_{\sigma}^{(k)}(x_1 | x_{-k+1}^0) - \\
 & - p_{\delta}(x_1 | x^0)| + |p_{\sigma}(x_1 | x^0) - p_{\delta}(x_1 | x_{-k+1}^0)| + \\
 & + |p_{\delta}(x_1 | x^0) - p_{\delta}^{(k)}(x_1 | x_{-k+1}^0)| \leq 2\varepsilon_{k-m-1} + d\eta_1.
 \end{aligned}$$

Используя лемму 14, получаем оценку четвертого слагаемого в правой части неравенства (4)

$$\begin{aligned}
 \| p_{\sigma}^{(k)}(x_{\infty} | x_{-k+1}^0) - p_{\delta}^{(k)}(x_{\infty} | x_{-k+1}^0) \| & \leq \\
 & \leq \frac{d\eta_1 + \varepsilon_{k-m-1}}{\delta_1^{k-1}} k \delta_2^{k-1} + \frac{\eta_2 (2k-1) \delta_2^{2k} \mu^{k-1}}{\delta_1^k}.
 \end{aligned}$$

Используя далее результаты лемм 8, 15 и 16, находим соответственно оценки первого и седьмого, второго и шестого, третьего и пятого слагаемых в правой части (4).

Окончательно имеем

$$\begin{aligned}
 \| p_{\sigma}(x_{\infty} | x^0) - p_{\delta}(x_{\infty} | x^0) \|_{L_1} & \leq 2\mu\varphi_m(t) + \\
 & + 2\delta_2\mu^2\varepsilon_{k-m} + 2\mu c[1 - (\mu\delta_1)^k]^{t/k} + \\
 & + \frac{d\eta_1 + \varepsilon_{k-m-1}}{\delta_1^{k-1}} k \delta_2^{k-1} + \frac{\eta_2 (2k-1) \delta_2^{2k} \mu^{k-1}}{\delta_1^k}. \quad (5)
 \end{aligned}$$

Покажем, что специальным выбором величин t, k, η_1 и η_2 можно сделать $\| p_{\sigma} - p_{\delta} \|_{L_1}$ сколь угодно малой. Условие

$$\lim_{k \rightarrow \infty} \frac{k_2 \delta_2^k}{\delta_1^k} \varepsilon_k = 0 \quad (6)$$

является достаточным для этого.

Рассмотрим сумму второго и третьего слагаемых в (5):

$$n(k, t) = 2\delta_2 \mu^2 e_{k-m} + 2c\mu [1 - (\mu\delta_1)^k]^{t/k}.$$

Эта функция достигает максимума по t в точке $t_{\min}^{(k)}$. Используя условие (6), можно показать, что выполняются равенства

$$\lim_{k \rightarrow \infty} \min_t n(k, t) = 0, \quad \lim_{k \rightarrow \infty} t_{\min}^{(k)} = \infty.$$

Таким образом, при выполнении условия (6) сумму $n(k, t)$ можно сделать сколь угодно малой, полагая $t = t_{\min}(k)$, а k — достаточно большим. Кроме того, первое слагаемое можно одновременно с $n(k, t)$ сделать сколь угодно малым.

Слагаемое $k \frac{\varepsilon_{k-m-1}}{\delta_1^{k-1}} \delta_2^{k-1}$ также стремится к 0, когда $k \rightarrow \infty$. Окончательно, подбирая η_1 и η_2 , можно сделать оставшиеся два слагаемых меньше любого наперед заданного числа. Тем самым установлено, что

$$\lim_{l \rightarrow \infty} \|p_\sigma(x_\infty | x^0) - p_\sigma(x_\infty | x^0)\|_{L_1} = 0.$$

Для доказательства утверждения теоремы заметим, что

$$\begin{aligned} |W_\sigma - W_\delta| &= \left| \int_X f(x_\infty) (p_\sigma(x_\infty | x^0) - p_\delta(x_\infty | x^0)) dx_\infty \right| \leqslant \\ &\leqslant \int_X |f(x_\infty)| |p_\sigma(x_\infty | x^0) - p_\delta(x_\infty | x^0)| dx_\infty \leqslant F \|p_\sigma - p_\delta\|_{L_1}. \end{aligned}$$

Значит, $\lim_{l \rightarrow \infty} W(\tilde{\sigma}) = W(\sigma)$. Поэтому из соотношения $|W_m - W_m^l| \leqslant |W_m - W(\sigma)| + |W(\sigma) - W(\tilde{\sigma})|$ следует, что $\lim_{l \rightarrow \infty} |W_m - W_m^l| \leqslant |W_m - W(\sigma)|$ для всех $\sigma \in D_m$. Следовательно, $\lim_{l \rightarrow \infty} W_m^l = W_m$. Теорема доказана.

Займемся адаптивными системами для классов управляемых стационарных процессов. В качестве цели управления назначим максимизацию предела (в среднем) среднего арифметического дохода, т. е. выполнения равенства

$$\text{l. i. m.}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T f(\xi_t) = \bar{W}.$$

Интересующие нас адаптивные системы реализуют следующую процедуру. На m -м шаге перебираются правила σ из множества D_m , имеющие одну и ту же глубину m . Каждое правило прилагается к процессу n_m тактов, и по откликам процесса строятся случайные величины

$$z_\sigma = \frac{1}{n_m} \sum_{t=t'-1}^{t'+n_m} f(\xi_t),$$

где t' — момент первого приложения правила σ . После завершения перебора элементов множества D_m в течение N_m тактов используется то правило, которое привело к максимальному z_σ . Затем переходим к $(m+1)$ -му шагу.

Через $S(\epsilon, \delta_1, \delta_2)$ обозначим класс управляемых стационарных процессов, у которых параметры $\epsilon'_1, \epsilon'_2, \dots$ подчинены неравенствам

$$\epsilon_k \geq \epsilon'_k, \quad k \geq 1,$$

а плотности условных вероятностей — неравенствам

$$0 < \delta_1 \leq p(x_1 | x^0, y^0) \leq \delta_2 < \infty,$$

причем выполнены условия (2) и

$$\lim_{k \rightarrow \infty} k \epsilon_k \left(\frac{\delta_2}{\delta_1} \right)^k = 0.$$

Теорема 2. *Определенная выше процедура обеспечивает в классе $S(\epsilon, \delta_1, \delta_2)$ цель управления (т. е. является адаптивной системой)*

$$\text{l. i. m. } \frac{1}{T} \sum_{t=1}^T f(\xi_t) = \bar{W},$$

если параметры ее выбраны из условий

$$\lim_{m \rightarrow \infty} |D_m^{k(m)}| \psi_m(n_m) = 0, \quad \lim_{m \rightarrow \infty} \frac{|D_m^{k(m)}| n_m}{N_{m-1}} = 0,$$

$k(m)$ — произвольная подпоследовательность натурального ряда, заданная заранее.

Доказательство. Введем обычным образом норму случайной величины $\|\eta\| = \sqrt{\mathbb{E}(\eta^2 | x^0, y^0)}$. Значит, $\mathbb{E} \left[\left(\frac{1}{T} \sum_{i=1}^T f(\xi_i) - \bar{W} \right)^2 | x^0, y^0 \right] = \left\| \frac{1}{T} \sum_{i=1}^T f(x_i) - \bar{W} \right\|^2$. Положим $a_m = \sum_{i=1}^m (|D_i^{k(i)}| n_i + N_i)$, $b_m = a_{m-1} + |D_m^{k(m)}| n_m$, т. е. a_m — момент окончания m -го шага адаптивной процедуры, а b_m — момент завершения сравнения правил с глубины m . Оценим норму

$$\begin{aligned} TI = & \left\| \sum_{i=1}^T (f(x_i) - \bar{W}) \right\| \leqslant \sum_{m=1}^l \left\| \sum_{i=b_m+1}^{a_m} (f(x_i) - \bar{W}) \right\| + \\ & + \sum_{m=1}^l \left\| \sum_{i=a_{m-1}+1}^{b_m} (f(x_i) - \bar{W}) \right\| + \\ & + \left\| \sum_{i=a_l+1}^{\min(T, b_{l+1})} (f(x_i) - \bar{W}) \right\| + \left\| \sum_{i=b_l+1}^T (f(x_i) - \bar{W}) \right\|. \end{aligned} \quad (7)$$

Здесь l — максимальное целое число такое, что $a_l \leqslant T$.

Займемся оценкой каждого из слагаемого в правой части (7). Сначала заметим, что

$$\begin{aligned} c_{1,m} = & \left\| \sum_{i=b_m+1}^{a_m} (f(x_i) - \bar{W}) \right\| \leqslant \left\| \sum_{i=b_m+1}^{a_m} (f(x_i) - W(\tilde{\sigma}_m)) \right\| + \\ & + \left\| \sum_{i=b_m+1}^{a_m} (W(\tilde{\sigma}_m) - W(\sigma_m^{k(m)})) \right\| + \left\| \sum_{i=b_m+1}^{a_m} (W(\sigma_m^{k(m)}) - \bar{W}) \right\|. \end{aligned}$$

Через $\tilde{\sigma}_m$ по-прежнему обозначается то случайное правило глубины m , которому отвечает наибольшая величина $z_\sigma = \frac{1}{n_m} \sum_{i=t'+1}^{t'+n_m-1} f(x_i)$. Поэтому $W(\tilde{\sigma}_m)$ неизменно в отрезке времени $[b_m + 1, a_m]$. Значит,

$$\begin{aligned} c_{1,m} \leqslant & \left\| \sum_{i=b_m+1}^{a_m} (f(x_i) - W(\tilde{\sigma}_m)) \right\| + \\ & + N_m \|W(\tilde{\sigma}_m) - W(\sigma_m^{k(m)})\| + N_m |W(\sigma_m^{k(m)}) - \bar{W}|. \end{aligned}$$

Согласно предыдущим результатам при любой предыстории

$$\mathbf{E} \left[\left(\frac{1}{N_m} \sum_{i=b_m+1}^{a_m} (f(x_i) - W(\tilde{\sigma}_m)) \right)^2 | x^{b_m}, y^{b_m} \right] \leq \psi_m(N_m)$$

или

$$\left\| \sum_{i=b_m+1}^{a_m} (f(x_i) - W(\tilde{\sigma}_m)) \right\| \leq N_m \sqrt{\psi_m(N_m)}.$$

Используя лемму 11, приходим к

$$c_{1,m} \leq N_m \sqrt{\psi_m(N_m)} + 2N_m \sqrt{|D_m^{k(m)}| \psi_m(n_m)} + \\ + N_m |W(\sigma_m^{k(m)}) - \bar{W}|.$$

Для слагаемых других типов в (7) без особого труда получаем оценки

$$c_{2,m} \leq 2F |D_m^{k(m)}| n_m, \\ c_3 \leq \left\| \sum_{i=a_l+1}^{\min(T, b_{l+1})} (f(x_i) - \bar{W}) \right\| \leq 2F |D_{l+1}^{k(l+1)}| n_{l+1}, \\ c_4 = \left\| \sum_{i=b_l+1}^t (f(x_i) - \bar{W}) \right\| \leq \\ \leq 2(T - b_{l+1}) \sqrt{|D_{l+1}^{k(l+1)}| \psi_{l+1}(n_{l+1})} + \\ + (T - b_{l+1}) \sqrt{\psi_{l+1}(T - b_{l+1})} + (T - b_{l+1}) |W(\sigma_{l+1}^{k(l+1)}) - \bar{W}|.$$

Заметим, что при $T < b_{l+1}$ имеем $c_4 = 0$. Подставим полученные оценки в (7). Тогда

$$I \leq \frac{1}{T} \left\{ \sum_{m=1}^l (N_m \sqrt{\psi(N_m)} + 2N_m \sqrt{|D_m^{k(m)}| \psi_m(n_m)} + \right. \\ \left. + N_m |W(\sigma_m^{k(m)}) - \bar{W}|) + 2F \sum_{m=1}^{l+1} |D_m^{k(m)}| n_m + \right. \\ \left. + \gamma [(T - b_{l+1})(\sqrt{\psi_{l+1}(T - b_{l+1})} + 2\sqrt{|D_{l+1}^{k(l+1)}| \psi_{l+1}(n_{l+1})} + \right. \\ \left. + |W(\sigma_{l+1}^{k(l+1)}) - \bar{W}|)] \right\},$$

где принято такое обозначение:

$$\gamma = \begin{cases} 1 & \text{при } b_{l+1} \leq T, \\ 0 & \text{при } b_{l+1} > T. \end{cases}$$

Из условий доказываемой теоремы, неравенства $T > \sum_{m=1}^l N_m$ и того, что для всякой положительной последовательности α_s такой, что $\sum_{s=1}^n \alpha_s \rightarrow \infty$, и стремящейся к 0 последовательности β_s справедливо равенство

$$\lim_{n \rightarrow \infty} \frac{\sum_{s=1}^n \alpha_s \beta_s}{\sum_{s=1}^n \alpha_s} = 0,$$

находим

$$\lim_{l \rightarrow \infty} \frac{1}{T} \left[\sum_{m=1}^l (N_m \sqrt{\psi_m(N_m)} + 2N_m \sqrt{|D_m^{k(m)}| \psi_m(n_m)} + + N_m |W(\sigma_m^{k(m)}) - \bar{W}|) + \sum_{m=1}^{l+1} 2F |D_m^{k(m)}| n_m \right] = 0. \quad (8)$$

В сделанном выводе принято также во внимание следующее равенство: $\lim_{m \rightarrow \infty} W(\sigma_m^{k(m)}) = \bar{W}$, в котором $W(\sigma_m^{k(m)}) = \sup_{\sigma \in D_m^{k(m)}} W(\sigma)$. Для установления этого равенства фиксируем целое число r и при всех $m \geq r$ рассмотрим величину $|\bar{W} - W(\sigma_m^{k(m)})|$. Заметим, что $W(\sigma_m^{k(m)}) \geq W(\sigma_r^{k(m)})$ и $W(\sigma_m^{k(m)}) \leq \bar{W}$, т. е.

$$\bar{W} - W(\sigma_m^{k(m)}) \leq |\bar{W} - W(\sigma_r^{k(m)})|.$$

Согласно теореме 1 $\lim_{k(m) \rightarrow \infty} W(\sigma_r^{k(m)}) = W(\sigma_r^{k(m)})$ и, следовательно,

$$\lim_{m \rightarrow \infty} |\bar{W} - W(\sigma_m^{k(m)})| \leq |\bar{W} - W_r|$$

для любого r . Принимая во внимание равенство $\bar{W} = \lim_{r \rightarrow \infty} W_r$, приходим к требуемому факту.

Из стремления $T \rightarrow \infty$ вытекает $l \rightarrow \infty$, поэтому равенство, аналогичное (8), справедливо и для предела при $T \rightarrow \infty$.

Нетрудно усмотреть далее, что

$$\begin{aligned} \lim_{T \rightarrow \infty} \gamma^2 \frac{T - b_{l+1}}{T} |D_{l+1}^{k(l+1)}| \psi_{l+1}(\eta_{l+1}) &= \\ &= \lim_{T \rightarrow \infty} |D_{l+1}^{k(l+1)}| \psi_{l+1}(n_{l+1}) = 0, \end{aligned} \quad (9)$$

а также

$$\lim_{T \rightarrow \infty} \gamma \frac{T - b_{l+1}}{T} |W(\sigma_{l+1}) - \bar{W}| = \lim_{l \rightarrow \infty} |W(\sigma_{l+1}) - \bar{W}| = 0. \quad (10)$$

Из того, что $\psi_{l+1}(n) \leq x$, при некотором x для любого n имеем

$$\begin{aligned} \gamma \frac{T - b_{l+1}}{T} \sqrt{\psi_{l+1}(T - b_{l+1})} &\leq \\ &\leq \begin{cases} \frac{n_{l+1}}{N_l} x, & \text{если } T - b_{l+1} < n_{l+1}, \\ \sqrt{\psi_{l+1}(n_{l+1})}, & \text{если } T - b_{l+1} \geq n_{l+1}. \end{cases} \end{aligned}$$

Из этого неравенства и условий теоремы можно вывести также

$$\lim_{T \rightarrow \infty} \gamma \frac{T - b_{l+1}}{T} \sqrt{\psi_{l+1}(T - b_{l+1})} = 0.$$

Отсюда, а также из (8)–(10) вместе с неравенством для I приходим к утверждению теоремы.

Полезно отметить два обстоятельства, относящиеся к изученной адаптивной системе:

- 1) реализуемая ею стратегия нестационарна,
- 2) не решается задача идентификации, т. е. оценки условных распределений, определяющих эволюцию процесса.

Обратимся теперь к следующей цели управления классом управляемых стационарных процессов: обеспечить

для любого процесса из рассматриваемого класса выполнение равенства

$$\lim_{T \rightarrow \infty} E(f(x_T) | x^0, y^0) = \bar{W}.$$

Конструкция соответствующей системы управления напоминает конструкцию предыдущей системы. На m -м шаге используются правила глубины m из множества D_m . Каждое прилагается к процессу n_m тактов, и то из них, которое максимизирует величину z_s , повторяется в продолжении случайного числа v_m тактов, а именно, с вероятностью p_m применение «наилучшего» правила на каждом такте независимо от предыстории может прекратиться, и тогда начинается $(m+1)$ -й шаг.

Теорема 3. Описанная выше процедура обеспечивает в классе $S(\epsilon, \delta_1, \delta_2)$ цель управления (т. е. является адаптивной системой)

$$\lim_{T \rightarrow \infty} E(f(x_T) | x^0, y^0) = \bar{W},$$

если параметры ее выбраны из условий

$$\lim |D_m^{k(m)}| \phi_m(n_m) = 0,$$

$$\sum_{m=1}^{\infty} p_1 p_2 \cdots p_{m-1} |D_m^{k(m)}| n_m < \infty.$$

Здесь снова $k(m)$ — произвольная подпоследовательность натурального ряда, задаваемая заранее.

Доказательство близко в идейном плане к доказательству теоремы 2 и ввиду его громоздкости опускается.

Обратим внимание на тот факт, что стратегия, к которой относится теорема 2, не может обеспечить назначенную здесь цель управления.

§ 2. α-однородные процессы

Здесь мы исследуем адаптивные системы для одного типа управляемых случайных процессов общего вида, т. е. характеризуемых совокупностью управляемых условных вероятностей $\mu_{t+1}(\cdot | x^t, y^t)$, в условиях которых фигури-

рует вся предыстория процесса (его значения и приложенные управлении). В качестве фазового пространства X этих процессов изберем конечный отрезок на числовой прямой (тем гарантируется существование математических ожиданий), без ограничения общности можно принять $X = [0, A]$.

Введем классы $\mathfrak{A}_{\alpha, r}$, управляемых случайных процессов, зависящие от двух параметров и называемые (α, r) -однородными процессами. Для этого через $y_r = \{y_{t_1}, \dots, y_{t_r}\}$ обозначим набор управлений длины r .

(α, r) -однородные процессы характеризуются требованием на условные математические ожидания: для данного $\alpha > 0$ и целого $r \geq 1$ при любых $t_1, t_2 \geq r$ и любого набора y_r справедливо неравенство

$$|\mathbb{E}(\xi_{t_1} | y_{t_1-r}, \dots, y_{t_1}) - \mathbb{E}(\xi_{t_2} | y_{t_2-r}, \dots, y_{t_2})| \leq \alpha,$$

где $y_{t_1-i} = y_{t_2-i} = y_i$ — i -е элементы набора y_r . В записи этого неравенства и всюду в настоящем параграфе подразумевается, что отсутствующие в условиях математических ожиданий прошлые управление и значения процесса являются произвольными.

Все процессы с фазовым пространством $[0, A]$ являются (A, r) -однородными при любом r . Впредь всюду будем подразумевать, что число α мало по сравнению с A .

Некоторые взаимосвязи введенных классов указаны в следующих включениях:

$$\mathfrak{A}_{\alpha, r} \supset \mathfrak{A}_{\beta, r}, \quad \alpha > \beta,$$

$$\mathfrak{A}_{\alpha, r_1} \subset \mathfrak{A}_{\alpha, r_2}, \quad r_1 > r_2.$$

К процессам из этих классов относятся обобщенные ПНЗ (§ 3 гл. VII).

Простейшими процессами являются П-процессы (§ 5 гл. VII). Таким образом, процессы из классов $\mathfrak{A}_{\alpha, r}$ представляют интерес. В случаях, когда значение числа r не играет роли, будем именовать (α, r) -однородные процессы короче — α -однородными.

Укажем еще два примера α -однородных процессов.

Пусть η_t — обобщенный в узком смысле ПНЗ, т. е. процесс, задаваемый мерой $\mu(\cdot | y_{t-r}^{t-1})$ (см. § 1 гл. VII),

а n_t — помеха, которая в каждый момент времени может зависеть от предыстории η_t и своей и, сверх того, подчиняется ограничению $|\eta_t| < \alpha/2$. Тогда управляемый случайный процесс $\xi_t = \eta_t + n_t$ α -однороден. Его реальный смысл — результат наблюдений за управляемым процессом, искаженный аддитивным коррелированным шумом. Если η_t — обобщенный ПНЗ, то $\xi_t = \eta_t + n_t$ также α -однородный процесс.

Следующий пример более содержателен. Обозначим через ξ_t количество полуфабриката, хранящегося на складах производства, перерабатывающего его в конечный продукт. В единицу времени расходуется η_t этого запаса. Производство организовано так, что за время r весь запас расходуется, а точнее,

$$E\xi_1 = E(\eta_t + \eta_{t+1} + \dots + \eta_{t+r-1}).$$

Введем управления, ими являются количества y_1, \dots, y_k заказываемого полуфабриката (пространство управлений Y конечно и состоит из k элементов). Через предыдущие значения введенных величин процесс выражается следующим образом:

$$\xi_t = \xi_{t-1} - \eta_{t-1} + y_{t-1}.$$

Выписывая последовательно значения ξ_{t-i} , приходим к такому представлению процесса

$$\xi_t = \xi_{t-r} - \eta_{t-1} - \eta_{t-2} - \dots - \eta_{t-r} + y_{t-1} + \dots + y_{t-r},$$

из которого в силу равенства

$$E(\xi_t | y_{t-r}^{t-1}) = y_{t-1} + \dots + y_{t-r}$$

выводим, что процесс ξ_t , описывающий производство продукта в описанных условиях, является $(0, r)$ -однородным.

Изложенное здесь дает основание рассмотреть класс \mathfrak{U} однородных процессов, который образован из процессов, обладающих свойством: при любых $t_1, t_2 \geq r$

$$|E(\xi_{t_1} | y_{t_1-1}, \dots, y_{t_1-r}) - E(\xi_{t_2} | y_{t_2-1}, \dots, y_{t_2-r})| \leq \alpha(r),$$

где $\alpha(r)$ — монотонно стремящаяся к нулю последовательность положительных чисел. Как и для (α, r) -однородных

процессов, мы считаем, что фазовым пространством служит отрезок $[0, A]$.

При управлении однородными и α -однородными процессами представляется естественным не ограничиваться лишь программными стратегиями y^∞ . Распределения этих процессов $\mu(\cdot | \xi^t, y^t)$ явным образом зависят от всех прошлых значений процесса и управлений. Поэтому разумно полагать, что стратегии общего вида (т. е. порожденные условными распределениями $F_t(N^t | \xi^{t-1})$, $t \geq 1$) способны расширить класс достижимых целей управления. Применимально к цели — максимизации величины

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t$$

— сравнение эффективности произвольных и программных стратегий содержится в теореме, относящейся к случаю, когда Y конечно.

Теорема 1. Пусть ξ_t — α -однородный процесс с неотрицательными значениями и конечным множеством $Y = \{y_1, \dots, y_k\}$. Тогда

$$\left| \sup_{\sigma} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - \sup_{y^\infty} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t \right| \leq \alpha.$$

Доказательство. Рассмотрим какую-нибудь непрограммную стратегию, образованную множеством $\{f_i\}$ условных распределений на Y . Как и ранее, через $E_\circ \xi_t$ обозначим математическое ожидание процесса по порожденной этой стратегией мере. Его можно записать в следующей форме (см. §§ 3, 4 гл. I):

$$\begin{aligned} E\xi_t &= \int_{X^{t-1} \times Y^{t-1}} \int_X x \mu_t(dx | x^{t-1}, y^{t-1}) \times \\ &\quad \times \prod_{i=1}^{t-1} \mu_i(dx_i | x^{i-1}, y^{i-1}) f_i(dy_i | x^{i-1}) = \\ &= \int_{Y^{t-1}} W(y_1, \dots, y_{t-1}) P(dy_1, \dots, dy_{t-1}), \end{aligned}$$

где $P(\cdot)$ — мера на конечном (по предположению) множестве наборов управлений длины $t-1$. Согласно опре-

делению α -однородного процесса найдется такое целое число r , что

$$\mathbf{E}(\xi_t | y^{t-1}) \leq \mathbf{E}(\xi_r | y_{t-r}, \dots, y_{t-1}) + \alpha.$$

Значит, используя обозначение $H(y_{t-r}^{t-1}) = \mathbf{E}(\xi_r | y_{t-r}^{t-1})$, получаем неравенство

$$\frac{1}{T} \sum_{t=r}^T \mathbf{E}_\sigma \xi_t \leq \mathbf{E}_\sigma \frac{1}{T} \sum_{t=r}^T H(y_{t-r}^{t-1}) + \alpha.$$

Будем интерпретировать последовательность значений неотрицательной функции $H(y_{t-r}^{t-1})$ как обобщенный в узком смысле ПНЗ, но детерминированный, и поэтому его значения принадлежат r -мерной матрице $M(k, r)$, на которой существует замкнутый путь длины $l_0(k, r)$ такой, что взятая по любому пути длины T сумма удовлетворяет неравенству (см. (1) в § 2 гл. VII)

$$\frac{1}{T} \sum_1^T H(y_{t-r}^{t-1}) \leq \left(1 + \frac{r-1}{T}\right) \max f(y_{t_0}, r).$$

Здесь $f(y_{t_0}, r)$ — среднее арифметическое элементов $M(k, r)$ по пути y_{t_0} . Пользуясь этой оценкой, находим

$$\overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E}_\sigma \xi_t \leq f(y_{t_0}, r) + \alpha.$$

В силу произвольности стратегии σ

$$\sup_\sigma \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E} \xi_t \leq f(y_{t_0}, r) + \alpha.$$

Использование программного управления y^ω в виде неограниченного повторения «оптимального» набора y_{t_0} приводит к

$$\left| \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbf{E} \xi_t - f(y_{t_0}, r) \right| \leq \alpha.$$

Добавляя к обоим последним неравенствам очевидное

$$\sup_{\sigma} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t \geq \sup_{y^\infty} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t,$$

убеждаемся в справедливости теоремы.

Теорема 2. Для α -однородных процессов ξ_t с неотрицательными значениями и конечным множеством $Y = \{y_1, \dots, y_k\}$ выполнено неравенство

$$\left| \sup_{y^\infty} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - \max f(y_{l_0(k,r)}, r) \right| < \alpha.$$

Доказательство протекает так же, как и доказательство теоремы в § 3 гл. VII.

Рассмотрим класс $\mathfrak{A}_{\alpha, r}(\alpha, r)$ -однородных процессов со значениями на отрезке $[0, A]$ и конечным пространством управлений. В качестве стратегий изберем модифицированные конечные автоматы $(CD)_{k,n}^{(l,r)}$ (см. § 4 гл. VII), которые ранее оказались ε -оптимальным семейством адаптивных систем для некоторых классов процессов. Сейчас мы хотим установить, обладают ли эти автоматы при управлении процессами из классов $\mathfrak{A}_{\alpha, r}$ способностью максимизировать (с точностью до заданного ε) величину

$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t$. Мы увидим, что автоматы тем точнее достигают максимума, чем меньше параметр α класса процессов.

Символ E здесь мы используем для обозначения математического ожидания по мере, которая порождена реализуемой автоматом $(CD)_{k,n}^{(l,r)}$ стратегий. Введем предельное математическое ожидание среднего циклического выигрыша (порядка r и длины l)

$$W(n; r, l) = \lim_{l \rightarrow \infty} \frac{1}{l} \sum_{i=1}^l E\xi_{t+i},$$

где n — глубина памяти автомата $(CD)_{k,n}^{(l,r)}$. Можно показать, что для процессов рассматриваемого класса

$$W(n; r, l) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t.$$

Теорема 3. Для произвольного процесса из класса $\mathfrak{A}_{\alpha, r}$ и любого $\varepsilon > 0$ существует глубина памяти n_ε автомата $(CD)_{k,n}^{(l_0(k,r),r)}$ такая, что при всех $n \geq n_\varepsilon$

$$W(n; r, l_0(k, r)) \geq \sup_{\sigma} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - 5\alpha - \varepsilon.$$

Доказательство. Как и в доказательстве теоремы 1 § 5 гл. VII, представим автомат $(CD)_{k,n}^{(l_0(k,r),r)}$, управляющий процессом ξ_t из класса $\mathfrak{A}_{\alpha, r}$, как автомат $D_{k^{l_0}, n}$, управляющий Π -процессом $f(\mathbf{y}_{l_0(k,r)}, r)$. Применив лемму из § 5 гл. VII, получим, что для любого $\varepsilon > 0$ существует n_ε такое, что для всех $n > n_\varepsilon$

$$W(n; r, l_0(k, r)) \geq \max f(\mathbf{y}_{l_0(k,r)}, r) - 3\alpha - \varepsilon.$$

Используя теоремы 1 и 2, получаем

$$W(n; r, l_0(k, r)) \geq \sup_{\sigma} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - 5\alpha - \varepsilon.$$

что и требовалось доказать.

Перейдем теперь к управлению однородными процессами класса \mathfrak{A} . Заметим сначала, что для этих процессов справедливы с очевидными корректировками теоремы 1 и 2. Зададимся целью управления — максимизировать $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t$, для процессов из \mathfrak{A} в качестве средства достижения цели изберем автоматы $(CD)_{k,n}^{l,r}$. Их свойство сформулировано в следующей теореме.

Теорема 4. Для произвольного процесса из класса \mathfrak{A} и любого $\varepsilon > 0$ существуют глубина памяти n_ε автомата

$(CD)_{k,n}^{(l_0(k,r),r)}$ и порядок r_ϵ такие, что при всех $n \geq n_\epsilon$ и $r \geq r_\epsilon$

$$W(n; r, l_0(k, r)) \geq \sup_{\sigma} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - \epsilon.$$

Доказательство. Зафиксируем процесс $\xi_t \in \mathfrak{A}$ и число $\epsilon > 0$. В качестве числа r_ϵ примем наименьшее значение r , для которого $\alpha(r) \leq \epsilon/6$. Справедливо включение $\xi_t \in \mathfrak{A}_{\epsilon/6, r_\epsilon}$ и по теореме 3 найдется такая глубина памяти n_ϵ , что при всех $n \geq n_\epsilon$ и $r \geq r_\epsilon$ выполнено неравенство

$$\begin{aligned} W(n; r, l_0(k, r)) &\geq \\ &\geq \sup_{\sigma} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - 5 \frac{\epsilon}{6} - \frac{\epsilon}{6} = \\ &= \sup_{\sigma} \overline{\lim}_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E\xi_t - \epsilon, \end{aligned}$$

которое доказывает теорему.

Таким образом, автоматы $(CD)_{k,n}^{(l,r)}$ образуют ϵ -оптимальное семейство адаптивных систем для класса \mathfrak{A} однородных управляемых процессов.

ГЛАВА X

ОБ АДАПТИВНОМ УПРАВЛЕНИИ ПРОЦЕССАМИ С НЕПРЕРЫВНЫМ ВРЕМЕНЕМ

Во введении было указано, что развитие теории адаптивных систем началось с построения и исследования систем управления для объектов, которые функционируют в непрерывном времени. Даже обзор накопленных в этом направлении фактов требует специальной монографии. Здесь мы ограничимся изучением лишь двух адаптивных систем для процессов с непрерывным временем.

Первая из систем предназначена для управляемых полумарковских процессов. Роль полумарковских процессов в проблемах массового обслуживания общеизвестна. Поэтому естественно быстрое развитие теории полумарковских процессов и возникновение концепции управляемых процессов такого рода.

Зададим конечное множество $S = \{1, \dots, N\}$ состояний и множество управлений $Y = \{y_1, \dots, y_k\}$. Пусть $\mathcal{P}^{(y)} = \|p_{ij}(y)\|$, $y \in Y$ — стохастические матрицы вероятностей переходов и $F = \{F_{ij}^{(y)}, i, j = 1, \dots, N\}$ — семейство распределений на положительной полуоси.

Управляемый полумарковский процесс (УПМП) задают системой $\{S, \mathcal{P}^{(y)}, F, Y\}$. Отличие их от марковских процессов — в распределениях времен пребывания в состояниях, которые у УПМП не обязательно экспоненциальные.

Состояние УПМП в момент t обозначим ξ_t . Траектории процесса предполагаются непрерывными справа. Условимся обозначать состояние после n -го перехода ($n \geq 0$) через ξ_n и через τ_n время пребывания в состоянии ξ_n до перехода в ξ_{n+1} . Будем считать управления y_t непрерывными справа: $y_{t+0} = y_t$. Если $\xi_{t-0} = \xi_t$, т. е. сохраняется состояние УПМП, то управление не изменяется.

Выскажем два предположения:

I. Для любых $y \in Y$, $i \in S$

$$0 < m_i(y) < \infty,$$

где $m_i(y) = \sum_{j=1}^N p_{ij}(y) \int_0^\infty x dF_{ij}(x | y)$ — среднее время пребывания в i -м состоянии, если выбрано управление y (суммирование ведется по тем j , для которых $p_{ij}(y) > 0$).

II. При любой марковской стратегии $y = (y(1), \dots, y(N))$ марковская цепь $(S, \|p_{ij}(y(i))\|)$ эргодическая. Соответствующее предельное распределение обозначим $\pi(y) = (\pi_1(y), \dots, \pi_N(y))$.

Сформулируем цель управления. Задана числовая матрица $R = \|r_{ij}\|$, $i \in S$, $j \in Y$. Определим случайную величину

$$\Psi_t(\sigma) = \frac{1}{t} \int_0^t r_{\xi_u y_u} du,$$

в обозначении которой фигурирует стратегия σ , определяющая избираемые действия. Требуется, чтобы при всех t , начиная с некоторого, эта величина не более чем на произвольное ϵ отличалась от величины

$$\sup_{\sigma} W(\sigma) = \bar{W} = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T E_{\sigma^0}(r_{\xi_u y_u}) du.$$

Верхняя грань в левой части берется по всем стратегиям, но можно показать, что достаточно ограничиться однородными марковскими стратегиями y . Тогда оказывается, что в предположении II справедливо равенство

$$\bar{W} = \frac{\sum_{i=1}^N \pi_i(y) m_i(y) r_{iy^0(i)}}{\sum_{i=1}^N \pi_i(y) m_i(y(i))}.$$

Разработаны способы достижения этой цели, т. е. синтеза стратегии σ^0 , в случае, когда известны матрицы $\mathcal{P}^{(y)}$ вероятностей переходов и $M = \|m_i(y)\|$ средних времен пребывания. В частности, полезен итерационный метод, сходный с изложенным в § 3 гл. VIII. Условимся обозначать так полученную оптимальную однородную марков-

скую стратегию $y^0 = y^0(\mathcal{P}^{(y)}, M)$. Положим, что она единственная.

Опишем конструкцию адаптивной системы для достижения сформулированной цели. Система является стохастической: правила выбора действий определяются набором вероятностей $b_{il}(n) = P(y_n = l | \xi_n = i)$, причем их начальные значения фиксированы:

$$b_{il}(0) = \frac{1}{k}, \quad i \in S, \quad l \in Y.$$

Трансформация этих вероятностей производится по результатам наблюдений за процессом. Введем необходимые обозначения.

Трехиндексная последовательность $N_{ijl}(n)$ указывает, сколько раз за n первых переходов происходил переход $(\xi_m = i) \rightarrow (\xi_{m+1} = j)$ при $y_m = l$, $m = 0, 1, \dots, n - 1$. Обозначим $N_{il}(n) = \sum_{j=1}^N N_{ijl}(n)$. Частоты таких переходов равны

$$p_{ijl}(n) = \begin{cases} 0, & \text{если } N_{il}(n) = 0, \\ \frac{N_{ijl}(n)}{N_{il}(n)}, & \text{если } N_{il}(n) > 0. \end{cases}$$

Матрицу эмпирических частот $p_{ijl}(n)$ обозначим $\mathcal{P}(n)$.

Эмпирические оценки средних времен пребывания в i -м состоянии (при выборе l -го управления) находятся по формулам

$$\mu_{il}(n) = \begin{cases} 0, & \text{если } N_{il}(n) = 0, \\ \frac{1}{N_{il}(n)} \sum_{m=0}^{n-1} \tau(\xi_m, \xi_{m+1}) \delta_{i, \xi_m} \delta_{l, y_m}, & \text{если } N_{il}(n) > 0, \end{cases}$$

где $\tau(\xi_m, \xi_{m+1})$ — время пребывания в состоянии ξ_m до перехода в состояние ξ_{m+1} . Совокупность чисел $\mu_{il}(n)$ обозначим $M(n)$.

По матрицам $\mathcal{P}(n)$ и $M(n)$ можно вычислить «псевдооптимальную» стратегию $y^0(\mathcal{P}(n), M(n))$,

Положим еще $v_i(n) = \min_l N_{il}(n)$ и пусть $\varepsilon(n)$ — числовая положительная последовательность такая, что

$$0 < \varepsilon(n) < 1, \quad \lim_{n \rightarrow \infty} \varepsilon(n) = 0.$$

Вероятности $b_{il}(n)$ преобразуются по таким правилам:

$$b_{il}(n+1) = \begin{cases} b_{il}(n) & \text{при } i \neq \xi_n, \\ 1/k & \text{при } i \neq \xi_n, v_i(n) = 0, \\ 1 - \varepsilon(v_i(n)) & \text{при } i = \xi_n, v_i(n) > 0, l = y_n^0(i), \\ \frac{\varepsilon(v_i(n))}{k-1} & \text{при } i = \xi_n, v_i(n) > 0, l \neq y_n^0(i). \end{cases}$$

В основе доказательства эффективности описанного алгоритма лежит следующий факт.

Лемма 1. Для всех $i, j \in S, l \in Y$, при которых $p_{ij}(y_l) > 0$, справедливо

$$\mathbf{P}(\lim_{n \rightarrow \infty} N_{ijl}(n) = \infty) = 1.$$

Доказательство утверждения происходит по тому же плану, что и аналогичного утверждения в § 3 гл. VIII (см. доказательство теоремы 2).

Следствие 1. В условиях леммы справедливо равенства

$$\mathbf{P}(\lim_{n \rightarrow \infty} p_{ijl}(n) = p_{ij}(y_l)) = 1,$$

$$\mathbf{P}(\lim_{n \rightarrow \infty} \mu_{il}(n) = m_i(y_l)) = 1.$$

Следствие 2. При всех $t > \tau(\omega)$, где $\mathbf{P}(\tau(\omega) < \infty) = 1$, $y_i^0 = y^0$ — оптимальная стратегия для УПМП.

Последний результат означает, что при всех i, l с вероятностью 1

$$\lim_{n \rightarrow \infty} b_{il}(n) = \begin{cases} 1, & l = y^0(i), \\ 0, & l \neq y^0(i). \end{cases}$$

При всякой однородной марковской стратегии σ с вероятностью единица имеем $\lim_{t \rightarrow \infty} \Psi_t(\sigma) = W(\sigma)$. Приняв на веру

этот «усиленный закон больших чисел», уже без труда заключаем:

Описанная система управления адаптивна в классе всех УПМП $\{S, \mathcal{P}^{(y)}, F^{(y)}, Y\}$ с одинаковыми S и Y , удовлетворяющими предположениям I и II.

Обратимся к другой задаче адаптивного управления процессами с непрерывным временем: к задаче стабилизации скалярного объекта, находящегося под воздействием случайных возмущений. Управляемый объект в отсутствие помех описывается линейным уравнением

$$\dot{x} = ax + by + f,$$

где a и b — постоянные, а $f=f(t)$ — измеримая ограниченная функция. Добавление помех приводит к стохастическому уравнению Ито

$$dx_t = (ax_t + by_t + f) dt + \\ + x_t q_1(t, x_t) dw_t^{(1)} + q_2(t, x_t) dw_t^{(2)}, \quad (1)$$

в котором $w_t^{(1)}$ и $w_t^{(2)}$ — независимые винеровские процессы, функции $q_i(t, x)$, $i=1, 2$, ограничены и измеримы. Они означают соответственно интенсивности параметрического и входного возмущений типа «белого шума».

Добавим к (1) законы управления

$$y = c(t)x, \quad (2)$$

$$dc = F(t, x, c) dt, \quad (3)$$

где $c(t)$ — числовая функция, а $F(t, x, c)$ удовлетворяет локальному условию Липшица по совокупности переменных (x, c) .

Зададим класс J объектов управления, характеризуемых уравнением (1), в котором

$$|q_i(t, x) - q_i(t, x')| \leq L_r |x - x'|, \quad i=1, 2, L_r > 0,$$

при любом $t \geq 0$ и $|x| < r$, $|x'| < r$. Коэффициенты уравнения a и b неизвестны, но знак b одинаков для всех элементов J и известен (либо плюс, либо минус). Можно показать, что при высказанных условиях система (1)–(3) имеет единственное решение — двумерный марковский процесс $z_t = (x_t, c_t)$, определенный до случайного

момента «обрыва» ζ . Этот процесс регулярен, если $P(\zeta = \infty) = 1$. Начальные условия (x_0, c_0) считаем произвольными. Через \mathcal{L} обозначим производящий оператор марковского процесса z_t .

Назначим цель управления. Любой процесс $z_t = (x_t, c_t)$ из класса J должен быть регулярным и

$$\limsup_{t \rightarrow \infty} E(x_t^2 + c_t^2) \leq D(z) < \infty,$$

т. е. требуется, чтобы этот предел был конечен.

Остается конкретизировать вид уравнений (3): примем

$$dc = -(\gamma x^2 + \alpha c) dt, \quad (4)$$

где $\alpha > 0$ и $\gamma \neq 0$.

Теорема 1. Система (1), (2), (4) обеспечивает в классе J заданную цель управления (т. е. является адаптивной), если $\gamma b > 0$ для всех элементов класса J .

Доказательство. Зафиксируем объект из J и введем функцию

$$V(x, c) = h_0 x^2 + (c - c^*)^2,$$

где c^* — число, $h_0 > 0$. Потребуем, чтобы функция V удовлетворяла неравенству

$$\mathcal{L}V(x, c) \leq -\kappa V(x, c) + \beta, \quad (5)$$

где $\kappa > 0$ и $\beta > 0$. Имеем

$$\begin{aligned} \mathcal{L}V(x, c) &= 2h_0 x (ax + bcx + f) - \\ &\quad - 2(c - c^*) (\gamma x^2 + \alpha c) + h_0 (q_1^2(t, x) x^2 + q_2^2(t, x)), \end{aligned}$$

откуда на основании неравенств $|f| \leq g_0$, $|q_i| \leq g_i$, $i = 1, 2$, получаем

$$\begin{aligned} \mathcal{L}V(x, c) &\leq 2h_0 (a + bc^*) x^2 + 2h_0 |x| g_0 - \\ &\quad - 2\alpha c (c - c^*) - 2(c - c^*) [\gamma x^2 - h_0 b x^2] + \\ &\quad + h_0 g_1^2 x^2 + h_0 g_2^2. \end{aligned}$$

Отсюда мы приедем к неравенству (5), если воспользуемся неравенствами

$$2g_0|x| \leq x^2 + g_0^2,$$

$$-2\alpha c|c - c^*| \leq -\kappa(c - c^*)^2 + \frac{\alpha^2(c^*)^2}{2\alpha - \kappa},$$

где принятые обозначения

$$h_0 = \frac{\gamma}{b}, \quad \beta = h_0 g_0^2 + h_0 g_2^2 + \frac{\alpha^2(c^*)^2}{2\alpha - \kappa},$$

фиксировано число $\kappa \in (0, 2\alpha)$ и число c^* выбрано из условия

$$2(a + bc^* + g_1^2) + 1 = -\kappa.$$

Покажем, что доказываемое утверждение вытекает из следующей леммы.

Лемма 2. Задано стохастическое уравнение Ито

$$dz_t = A(t, z_t) dt + Q_1(t, z_t) dw_t^{(1)} + Q_2(t, z_t) dw_t^{(2)}, \quad (6)$$

где $z_t \in R_n$, $w_t^{(1)}$, $w_t^{(2)}$ — l -мерные независимые винеровские процессы. Допустим, что вектор-функция $A(t, z_t)$ и матрицы-функции $Q_i(t, z)$ непрерывны в $(t \geq 0) \times R_n$, удовлетворяют локальному условию Липшица по z и

$$\|Q_1(t, z)\| \leq g_1 \|z\|, \quad \|Q_2(t, z)\| \leq g_2, \quad g_1, g_2 > 0.$$

Кроме того, пусть существует функция $V(z) > 0$, дважды непрерывно дифференцируемая и подчиненная неравенствам (для некоторых $\kappa_1, \kappa_2, \kappa_3, \alpha, \beta > 0$ и любых z)

$$\|z\|^2 \leq \kappa_1(1 + V(z)), \quad \|\nabla V(z)\|^2 \leq \kappa_2(1 + V(z)),$$

$$|\mathcal{L}V(z)| \leq \kappa_3(1 + V(z)), \quad \mathcal{L}V(z) \leq -\kappa V(z) + \beta,$$

где \mathcal{L} — производящий оператор, определяемый уравнением (6).

Тогда при любом (возможно случайном) начальном условии z_0 , не зависящем от процессов $w_t^{(1)}, w_t^{(2)}$, существует единственное с точностью до эквивалентности решение z_t уравнения Ито, которое является регулярным процессом

и при всех $t \geq 0$ удовлетворяет неравенству

$$\mathbf{E} V(z_t) \leq \frac{\beta}{\alpha} + \left[\mathbf{E} V(z_0) - \frac{\beta}{\alpha} \right] e^{-\alpha t},$$

если существует $\mathbf{E} V(z_0)$.

Доказательство леммы опускается.

В рассматриваемой нами ситуации условия этой леммы выполнены. Надо положить $z_t = (x_t, c_t)$. Значит,

$$\mathbf{E} \|z_t\|^2 \leq \kappa_1 (1 + \mathbf{E} V(z_t)) \leq \kappa_1 \left[1 + \frac{\beta}{\alpha} + \left[\mathbf{E} V(z_0) - \frac{\beta}{\alpha} \right] e^{-\alpha t} \right].$$

Отсюда сразу получаем

$$\overline{\lim_{t \rightarrow \infty}} \mathbf{E} (x_t^2 + c_t^2) \leq \kappa_1 + \beta \frac{\kappa_1}{\alpha} < \infty.$$

Теорема доказана.

Достигнутый результат может быть обобщен в разных направлениях: на многомерный случай, на помехи с конечным вторым моментом и т. п. Мы не станем здесь их обсуждать.

ПРИМЕЧАНИЯ

К главе I. Содержание этой главы традиционно. Подробное изложение затронутых здесь вопросов можно найти в руководствах по теории вероятностей (Колмогоров, Боровков, Дуб, Гихман и Скороход) и по теории управления. Об управляемых случайных процессах см. Дынкин.

Результаты о немарковском моменте последнего достижения суммой независимых случайных величин заданного уравнения получены Зигмундом, Роббинсом и Веделем.

К главе II. Понятие аддитивной системы в изложенном виде предложил Срагович [2], [5]. Первое формальное определение, но в более узких рамках, дал Якубович [2], [3], [4].

Обучаемые системы являются обобщением «стохастической модели» Буша и Мостеллера, которая была предметом многих тонких исследований (Иосифеску и Теодореску, Норман). Первоначально такие системы ввели Оническу и Михок в форме «систем с полными связями».

Способы фактической реализации автоматов рассматриваются во многих работах. О конечных автоматах см. Глушков, о вероятностных — Агасандян и Срагович, об автоматах с переменной структурой см. Агасандян и Вехлер. Кроме того, существует безгранична техническая литература.

Задачу оптимизации на конечном интервале с помощью байесовых оценок рассматривал Аоки, следуя, в частности, Беллману.

К главе III. Автоматы $D_{k,n}$ впервые предложил Роббинс (впоследствии они переоткрывались), вслед за которым подобными конструкциями занялись другие исследователи (например, Смит и Пайк). Систематическое изучение автоматов начал Цетлин. В его книге содержится дальнейшая библиография. Цетлин называл ОПНЗ «стационарной случайной средой», а ϵ -оптимальные автоматы «автоматами из асимптотически оптимальных последовательностей». Квазилинейные автоматы открывались дважды, сначала Канделаки и Церцвадзе, а затем Валахом. Автоматы со сравнивающей цепочкой также имеют несколько авторов, среди них Смит и Пайк, Клейменова и Мудров (см. также книгу Цетлина). Среднее время совершения действия ввел и использовал Роббинс.

Первым использовал конечные автоматы Роббинса — Цетлина для управления не бинарными ОПНЗ Гурвич. Несколько позднее так поступал Попов, в том числе для процессов более общего вида, чем ОПНЗ (см. гл. VII и IX).

Оценивающие автоматы определил Гурвич [3], который исследовал их свойства при управлении ПНЗ (однородными и неоднородными).

Примеры δ -автоматов появились в работах Сраговича и Флерова [1], [2]. Дальнейшее развитие их теории принадлежит Сраговичу [1], [3]. Оценки среднего времени обучения нашел Коновалов, теорема 3 (§ 6) публикуется впервые.

Лемма 1 (§ 5) принадлежит Хеффдингу. О вероятностях больших уклонений см. Петров.

Конструкция многих оптимальных автоматов ввел Флеров [2]. Изложеный здесь метод их исследования принадлежит Засухину [2]. Результат об автоматном поиске условного экстремума получен Поповым, публикуется впервые.

Задачу прогноза марковских процессов рассмотрел Холево. О приложениях к управлению запасами см. Сильвестрова.

К главе IV. Свойства конечных автоматов при управлении бинарными неоднородными ПНЗ первым изучал Цетлин, ими интересовались также другие исследователи. Результаты об оценивающих автоматах принадлежат Гурвичу [3].

Возможности δ -автоматов при управлении неоднородными ПНЗ рассматривал Срагович, который также ввел их модификации [3].

К главе V. Развитие метода стохастической аппроксимации началось с работы Роббинса и Монро, за которой последовала работа Кифера и Вольфовица. Их схемы обобщил Дворецкий. Впоследствии эта тематика необозримо разрослась. Математическую теорию этого метода изложили Невельсон и Хасьминский, см. также Шметтерер. Теорема 1 в § 2 доказана Гладышевым. Скорость сходимости процедуры Кифера — Вольфовица оценил Дупач. Полезный результат о моменте достижения ε -окрестности предела установили Зигмунд, Роббинс, Вендель, см. также Липцер, Ширяев (последний параграф).

На роль стохастической аппроксимации (и вообще рекуррентных процедур) для построения адаптивных систем разнообразного назначения первым обратил внимание Цыпкин [1], [2], там же содержится обширная библиография (см. также Цыпкин и Поляк). Эти методы развиваются (в частности, применительно к проблемам распознавания) в монографии Айзermana, Бравермана и Розеноэра.

Задачу прогноза стационарных процессов рассмотрел Юринский. Стохастическое программирование разрабатывается с 1955 г. Монографическое изложение дано Юдиным, Ермольевым. С иными подходами можно познакомиться у Беряну, где указана дополнительная литература.

К главе VI. Представление об играх автоматов развил Цетлин в интересах моделирования физиологических явлений. Вместе с сотрудниками он рассмотрел имитационными методами много примеров игр. Среди аналитических результатов отметим результаты Сушкича. Игру Гура определили Боровиков и Брызгалов.

Приложения игр автоматов (как децентрализованных систем управлений) изучали Бутрименко и Лазарев, Стефанюк и др.

Результаты §§ 2, 3 получил Гурвич [1], [2], [3].

Численные примеры в § 4 заимствованы из работ Цетлина с сотрудниками, работы Сраговича и Шапиро. Другие факты имеются у Иванова, Флерова [1], [3].

Иерархические системы определяются согласно Сраговичу [4]. Пример двухуровневой системы управления заимствован из статьи «Поведение автоматов в периодических случайных средах и задача синхронизации при наличии помех» в книге Цетлина (написана с сотрудниками), в которой оптимальные свойства этой системы проверялись экспериментально. Сформулированная теорема принадлежит Коновалову и публикуется впервые.

К главе VII. Результаты получены Поповым [3].

К главе VIII. Результаты § 2 принадлежат Попову [2]. Изложение в § 3 основано на статьях Засухина [1], [3], в которых также изучены аддитивные системы для «связных марковских цепей» с доходами (для каждой пары состояний переход требует применения своей однородной стратегии). Исследование систем типа *RZ* инициировано работой Риордона.

Итерационные алгоритмы отыскания оптимальной однородной марковской стратегии см. в книге Ховарда и у Мине и Осаки. Доказательство того, что оптимальная стратегия находится среди однородных марковских, первыми дали Висков и Ширяев, см. также Мине и Осаки, Дынкин и Юшкевич.

Концепцию конечно-сходящихся алгоритмов для решения целевых неравенств выдвинул Якубович [1]—[7]. Ему принадлежат приведенные алгоритмы. Аналогичный метод решения линейных уравнений был ранее предложен Качмажем. Теорему 1 (§ 4) доказал Фомин [1]. Задачу о стабилизации линейного объекта поставил (и первым дал решение) Якубович [6]. Приведенная здесь форма ее решения принадлежит Любаческому. Конечно-сходящиеся алгоритмы с успехом применяются в проблемах распознавания, идентификации (Фомин [2], Тимофеев). Они допускают распространение на стохастический вариант (Фомин [1]), на задачи с непрерывным временем (например, Фрадков).

К главе IX. Результаты § 1 получены Агасандяном [1], [2]. Результаты § 2 принадлежат Попову [3].

К главе X. Метод аддитивного управления полумарковскими процессами предложил Засухин. Стабилизацию стохастических объектов исследовал Фрадков, которому принадлежит впервые публикуемая теорема 1. Ее можно распространить на многомерный случай. Отметим работу Мандла об управлении скачкообразными марковскими процессами.

Мы не касаемся здесь процедур стохастической аппроксимации в непрерывном времени. Результаты содержаться у Невельсона и Хасьминского, Красулиной. Обзор других работ дал Юдин.

ЛИТЕРАТУРА

А г а с а н д р и н Г. А.

1. Адаптивное управление однородными случайными процессами, в сб. «Исследования по теории адаптивных систем», М., ВЦ АН СССР, 1976.
2. Адаптивная система для класса однородных последовательностей, ДАН СССР 228, 2 (1976).

А г а с а н д р и н Г. А., В е х л е р В.

Разложение конечного автомата на управляющий и управляемый подавтоматы, Известия АН СССР, Техническая кибернетика 4 (1976).

А г а с а н д р и н Г. А., С р а г о в и ч В. Г.

О структурном синтезе вероятностных автоматов, Известия АН СССР, Техническая кибернетика 6 (1971).

А й з е р м а н М. А., Б р а в е р м а н Э. М., Р о з о н о э р Л. И. Метод потенциальных функций в теории обучения машин, М., «Наука», 1970.

А о к и М.

Оптимизация стохастических систем, М., «Наука», 1971.

Б е л л м а н Р.

Процессы регулирования с адаптацией, М., «Наука», 1964.

Б е р я н у (В е г е а п и В.)

On stochastic linear programming IV, Proceedings of the 4 Conference on Probability Theory, Bucaresti, 1975.

Б о р о в и к о в В. А., Б р ы з г а л о в В. И.

Простейшая симметрическая игра многих автоматов, Автоматика и телемеханика 26, 4 (1965).

Б о р о в и к о в А. А.

Теория вероятностей, М., «Наука», 1976.

Б р ы з г а л о в В. И., П я т е ц к и й - Ш а п и р о И. И., Ш и к М. Л.

О двухуровневой модели взаимодействия автоматов, ДАН СССР 160, 5 (1965).

Б у т р и м е н к о А. В., Л а з а р е в В. Г.

Система поиска оптимальных путей передачи сообщений, Проблемы передачи информации 1, 1 (1965).

Б у ш Р., М о с т е л л е р Ф.

Стохастические модели обучаемости, М., «Наука», 1958.

В а л а х В. Я.

О поведении автомата с избирательной тактикой в стационарных случайных средах, Кибернетика 4 (1968).

- В иск о в О. В., Ши ря е в А. Н.
Об управлении, приводящих к стационарным режимам,
Тр. МИАН СССР 71 (1964).
- Г и х м а н И. И., Ско ро х од А. В.
Теория случайных процессов, М., «Наука», т. I, 1971; т. II,
1973; т. III, 1975.
- Г ла ды шев Е. Г.
О стохастической аппроксимации, Теория вероятностей и ее
применения 10, 2 (1965).
- Г лу шко в В. М.
Синтез цифровых автоматов, М., «Наука», 1962.
- Г ур ви ч Е. Т.
1. Метод асимптотического исследования игр автоматов, Автома-
тика и телемеханика 2 (1975).
2. Адаптивное управление векторными случайными процессами,
в сб. «Исследования по теории адаптивных систем», М., ВЦ АН
СССР, 1976.
3. Простейшая модель выбора действий в условиях неопреде-
лленности, Автоматика и телемеханика 11 (1976).
- Д вор ец кий (Двог ет ск у А.)
On stochastic approximation, Proceedings of the 3 Berkeley Sym-
posium 1 (1956).
- Д уб Д ж.
Вероятностные процессы, М., ИЛ, 1956.
- Д уп ач (Дир а ё V.)
Notes on stochastic approximation method, Czech. Math. Journ.
1 (1958).
- Дынкин Е. Б.
Управляемые случайные последовательности, Теория вероят-
ностей и ее применения 10, 1 (1965).
- Дынкин Е. Б., Юшкевич А. А.
Управляемые марковские процессы и их приложения, М.,
«Наука», 1975.
- Е рм ольев Ю. М.
Методы стохастического программирования, М., «Наука»,
1976.
- Засухин В. С.
1. Адаптивные системы управления марковскими цепями с до-
ходами, Известия АН СССР, Техническая кибернетика 5
(1973).
2. Достаточное условие асимптотической оптимальности одного
класса вероятностных автоматов, Известия АН СССР, Техни-
ческая кибернетика 5 (1975).
3. Адаптивное управление марковскими цепями с доходами,
в сб. «Исследования по теории адаптивных систем», М.,
ВЦ АН СССР, 1976.
- Зигмунд, Роббинс, Венделль (Sieg m und D.,
Rob bins H., Wendel J.)
The limiting distribution of the last time $S_n \geq n\varepsilon$, Proceedings
Nat. Acad. Sci. 61, 1 (1968).

И ван о в В. Н.

Поведение автоматов типа G в матричной игре против автоматов с линейной тактикой, в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1971.

И осифеску, Теодореску (Josifescu M., Teodorescu R.)

Random processes and learning, Springer-Verlag, Berlin, 1969.

Канделаки Н. П., Церцвадзе Г. Н.

О поведении некоторых классов стохастических автоматов в случайных средах, Автоматика и телемеханика 27, 6 (1966).

К ач ма ж (Kaczmarz S.)

Angenäherte Auflösung von Systemen Linearer Gleichungen, Bull. Internat. Acad. Polon. Sci. A., 1937.

К и ф е р, В ольфович (Kiefer E., Wolfowitz J.)

Stochastic estimation of the maximum of a regression function, Annals of Math. Stat. 23, 3 (1952).

К лей менова Г. С., М удр ов В. И.

Об одной конструкции автомата, асимптотически оптимального в стационарной случайной среде, Кибернетика 3 (1968).

К олмогоров А. Н.

Основные понятия теории вероятностей, М., «Наука», 1974.

К оновалов М. Г.

О характеристиках времени обучения автоматов типа G , в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1976.

К расулина Т. П.

О стохастической аппроксимации для случайных процессов с непрерывным временем, Теория вероятностей и ее применения 16, 4 (1971).

Л ипп цер Р. Ш., Ширяев А. Н.

Статистика случайных процессов, М., «Наука», 1974.

Л ю бачевский Б. Д.

Рекуррентный алгоритм адаптивного управления линейным динамическим объектом, Автоматика и телемеханика 4 (1974).

М андл (Mandl P.)

On the adaptive control of finite state Markov processes, Zeitschrift für Wahrscheinlichkeits-theorie und verwandte Gebiete 27, 4 (1973).

М ине, Осаки (Mine H., Osaki D.)

Markovian decision processes, New York, 1970.

Н евельсон М. Б., Х асьминский Р. З.

Стохастическая аппроксимация и рекуррентное оценивание, М., «Наука», 1972.

Н орман (Norman M. F.)

Mathematical learning theory, в кн. «Mathematics of the decision sciences», р. 2, 1968.

О ни ческу, Михок (Onicescu O., Mi h o c G.)

Sur les chaînes statistique, Compte r. Acad. Sci. 200, 1935.

П етров В. В.

Суммы независимых случайных величин, М., «Наука», 1972.

Попов Ю. В.

1. Синтез адаптивной системы для однородного случайного процесса, Известия АН СССР, Техническая кибернетика 5 (1973).
2. Адаптивное управление эргодическими марковскими цепями, в сб. «Исследования по теории адаптивных систем», М., ВЦ АН СССР, 1976.
3. Адаптивные системы управления некоторыми классами случайных процессов общего вида, в сб. «Исследования по теории адаптивных систем», М., ВЦ АН СССР, 1976.

Риордон (R i o r d o n S.)

An adaptive automation controller for discrete time Markov processes, Automatica 5 (1969).

Роббинс (R o b b i n s H.)

A sequential decision problem with a finite memory, Proc. Nat. Acad. Sci. USA 42, 3 (1956).

Роббинс, Монро (R o b b i n s H., M o n r o S.)

A stochastic approximation Method, Annals of Math. Statistics 22, 3 (1951).

Сильвестрова Э. М.

Об одной адаптивной системе управления запасами, Кибернетика 4 (1973).

Смит, Пайк (S m i t h C., P u k e R.)

The Robbins—Isbell two-armed-bandit problem with finite memory, Annals of Math. Stat. 36, 5 (1965).

Срагович В. Г.

1. Автоматы с многозначным входом и их поведение в случайных средах, в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1971.

2. Адаптивные управляющие системы и автоматы, Известия АН СССР, Техническая кибернетика 2 (1972).

3. Оптимальные свойства одного класса управляющих автоматов, Известия АН СССР, Техническая кибернетика 1 (1973).

4. К теории иерархических систем. Основные понятия, Известия АН СССР, Техническая кибернетика 2 (1975).

5. О понятии адаптивной системы, в сб. «Исследования по теории адаптивных систем», М., ВЦ АН СССР, 1976.

Срагович В. Г., Флеров Ю. А.

1. Построение класса оптимальных автоматов, ДАН СССР 159, 6 (1964).

2. Об одном классе стохастических автоматов, Известия АН СССР, Техническая кибернетика 4 (1965).

Срагович В. Г., Шапиро Л. З.

О коллективном поведении автоматов типа G , в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1971.

Степанюк В. Л.

Поведение коллектива автоматов в задаче о регулировке мощности, сб. «Проблемы кибернетики» 20 (1968).

Сушкин Б. Г.

О симметрических играх больших коллективов вероятностных автоматов, Проблемы передачи информации 9, 3 (1973).

Тимофеев А. В.

Рекуррентные конечно-сходящиеся алгоритмы идентификации дискретных динамических объектов, Известия АН СССР, Техническая кибернетика 6 (1973).

Феллер В.

Введение в теорию вероятностей и ее приложения, М., «Мир», т. 1, 1964; т. 2, 1967.

Флеров Ю. А.

1. Об играх стохастических автоматов, в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1967.

2. О некоторых классах многовходовых автоматов, в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1971.

3. Многоуровневые динамические игры, в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1971.

Фомин В. Н.

1. Стохастические аналоги конечно-сходящихся алгоритмов обучения опознающих систем, в сб. «Вычислительная техника и вопросы программирования», Л., изд.-во ЛГУ, вып. 6, 1971.

2. Математическая теория обучаемых опознающих систем, Л., изд.-во ЛГУ, 1976.

Фрадков А. Л.

Синтез адаптивной системы стабилизации линейного динамического объекта, Автоматика и телемеханика 12 (1974).

Хеффдинг (Hoeffding W.)

Probability inequalities for sum of bounded random variables, Journal of the Amer. Statist. Assoc. 58, 301 (1963).

Ховард Р.

Динамическое программирование и марковские процессы, М., «Сов. радио», 1964.

Холево А. С.

Автоматы, прогнозирующие случайный процесс, ДАН СССР 164, 3 (1965).

Цетлин М. Л.

Исследования по теории автоматов и моделирование биологических систем, М., «Наука», 1969.

Цыпкин Я. З.

1. Адаптация и обучение в автоматических системах, М., «Наука», 1968.

2. Алгоритмы функционирования адаптивных систем, в сб. «Вопросы кибернетики. Адаптивные системы», М., «Наука», 1974.

Цыпкин Я. З., Поляк Б. Т.

Достижимая точность алгоритмов адаптации, ДАН СССР, 218 3 (1974).

Шметтерер (Schmetterer L.)

L'approximation stochastique, Université de Clermont-Ferrand, 1972.

Юдин Д. Б.

Математические методы управления в условиях неполной информации, М., «Сов. радио», 1974.

Юринский В. В.

Метод стохастической аппроксимации в задаче прогноза стационарной последовательности, в сб. «Исследования по теории самонастраивающихся систем», М., ВЦ АН СССР, 1967.

Якубович В. А.

1. Рекуррентные конечно-сходящиеся алгоритмы решения системы неравенств, ДАН СССР 166, 6 (1966).
2. К теории адаптивных систем, ДАН СССР 182, 3 (1968).
3. Адаптивные системы с многошаговыми целевыми условиями, ДАН СССР 183, 2 (1968).
4. Об одной задаче обучения целесообразному поведению, Автоматика и телемеханика 8 (1969).
5. Конечно-сходящиеся алгоритмы решения систем неравенств и их применение в задачах синтеза адаптивных систем, ДАН СССР 189, 3 (1969).
6. Об одном методе построения адаптивного управления линейным динамическим объектом в условиях большой неопределенности, в сб. «Вопросы кибернетики. Адаптивные системы», М., «Наука», 1974.
7. Метод рекуррентных целевых неравенств в теории адаптивных систем, в сб. «Вопросы кибернетики. Адаптивное управление», М., «Наука», 1976.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

- Автомат асимптотически оптимальный 118
 - вероятностный Мура 44
 - «глубокий» $D_{k,n}$ 71
 - детерминированный 45
 - с переменной структурой 45
 - — тактикой гистерезисной 78
 - — — линейной L_k, n 73
 - со сравнивающей цепочкой 81
 - типа G 139
 - A_α 120
 - $A_{\delta\alpha}$ 120
 - $A_{\delta\beta}$ 133
 - $A_{\delta h}$ 135
 - FP 123
 - G 100
 - K_k, n 75
 - SA 121
 - Z 121
- α -автомат 118
- Автоматы $(CD)_{k,n}^{(l,r)}$ 215
 - $(MD)_{k,n}^{(l,m)}$ 233
 - $(Md)_{k,n}^{(l)}$ 238
- $\delta\omega$ -автоматы 89
- Адаптивная система 39
- Адаптивный прогноз 142
- Алгоритм конечно-сходящийся (КСА) 261
 - GM 243
 - $GM(n^0, \delta)$ 243
 - $GM(n^0, 0)$ 246
 - $GM(\delta)$ 248
 - RZ 250
- Асимптотическая оптимальность 35
- Ассоциативный марковский процесс 50
- Байесовский подход 57
- Бореля—Кантелли лемма 118
- Выигрыш (доход) 31
 - средний 31
- Градиент обобщенный 163
- Граф игры 178
- Гура игра 181
- Действие (управление) 25
 - максиминное 129
 - оптимальное 86
- Децентрализованная система управления 167
- Дини теорема 73
- Игра 167
 - автоматов 167
 - в размещения 169
 - на окружности 190
 - с общей кассой 181
- Исход партии 167
- Коэффициент забывания 133
- Марковская цепь 21
 - — с доходами 29
- Марковский момент 21
- Мартингал 24
- Матрица средних значений 206
- Метод идентификации 56
 - Монте-Карло 188
 - стохастической аппроксимации (МСА) 149
- Моделирование игр автоматов 188
- Модель Буша—Мостеллера 43

- Обобщенные ПНЗ 211
 — — в узком смысле 202
 — — — широком смысле 203
- Обучаемая система 42
 — — δ -оптимальная 86
- Однородные процессы 295
 — — с независимыми значениями (ОПНЗ) 28
- (α, r)-однородные процессы 294
- Одношаговая прибыль 239
- ε -оптимальное семейство аддитивных систем 47
- ε -оптимальность 34
- Партия 167
 — максиминная 168
- Платежная функция 167
- Полумартингал 24
- Правило выбора действий 26
- Предельный средний (одношаговый) доход 239
- Процедура Кифера—Вольфовича (ПКВ) 157
 — рекуррентная 148
 — Роббинса—Монро (ПРМ) 151
- π -процесс 218
- Процессы с независимыми значениями (ПНЗ) 28
- Система управления 17
- Ситуация равновесия 168
- Случайная величина 18
- Случайный процесс 19
- Среднее время обучения 100
 — — совершения действия 68
- Средний циклический выигрыш 205
- Статистика процесса 42
- Стохастический квазиградиент 163
- Стохастическое программирование 162
 — уравнение Ито 315
- Стратегия 26
 — марковская 27
 — оптимальная, итеративный метод вычисления 249
 — программная 27
 — стационарная 27
- Управляемая условная вероятность 25
- Управляемый процесс марковский 28
 — — полумарковский (УПМП) 301
 — — случайный 25
 — — стационарный 273
- Цель управления 29
- Циклический выигрыш 205
- Элементарная управляющая система 41
- Эргодическая марковская цепь 23

Владимир Григорьевич Срагович

Теория адаптивных систем

М, 1976 т., 320 стр. с илл.

Редакторы *Н. П. Рябенъкая, Г. А. Агасандян*

Тех. редактор *К. Ф. Брудно*

Корректор *Л. С. Сомова*

* * *

Сдано в набор 17/V 1976 г. Подписано к печати
3/VIII 1976 г. Формат бумаги 84×108/32 тип. № 1.
Физ. печ. л. 10. Усл. печ. л. 16,80. Уч.-изд. л.
15,47. Тираж 7000 экз. Т-15113. Цена 1 р. 18 к.
Тип. зак. № 1256

* * *

Издательство «Наука»
Главная редакция
физико-математической литературы
117071, Москва, В-71, Ленинский проспект, 15

* * *

1-я тип. издательства «Наука».
199034, Ленинград, В-34, 9 линия, д. 12

