УДК 004.85
ББК 32.971.3
С21

Саттон Р. С., Барто Э. Дж.
С21 Обучение с подкреплением: Введение / пер. с англ. А. А. Слинкина. – М.: ДМК Пресс, 2020. – 552 с.: ил.

ISBN 978-5-97060-097-9

Идея обучения с подкреплением возникла десятки лет назад, но этой дисциплине предстояло пройти долгий путь, прежде чем она стала одним из самых активных направлений исследований в области машинного обучения и нейронных сетей. Сегодня это предмет интереса ученых, занимающихся психологией, теорией управления, искусственным интеллектом и многими другими отраслями знаний.

Подход, принятый авторами книги, ставит акцент на практическое использования обучения с подкреплением. В первой части читатель знакомится с базовыми его аспектами. Во второй части представлены приближенные методы решения в условиях ограниченных вычислительных ресурсов. В третьей части книги обсуждается важность обучения с подкреплением для психологии и нейронаук.

Издание предназначено для студентов технических вузов, разработчиков, специализирующихся на машинном обучении и искусственном интеллекте, а также представителей нетехнических профессий, которые могут использовать описанные методики в своей работе.

УДК 004.85
ББК 32.971.3

ISBN 978-0-262-03924-6 (англ.)
ISBN 978-5-97060-097-9 (рус.)

Посвящается памяти А. Гарри Клопфа

Содержание

Вступительное слово от ГК «Цифра»

Прошло уже несколько лет с тех пор, как наша команда ступила на путь применения искусственного интеллекта для совершенствования процессов в промышленности и логистике. В самом начале мы и представить не могли, насколько тернистой, но в то же время невероятно интересной окажется эта дорога. За это время мы успели поработать с различными производствами и решить множество задач – от оптимизации производства битумных материалов до улучшения системы распределения нефтепродуктов и внедрения систем машинного зрения на карьерные экскаваторы. Методы машинного обучения, которые используются для решения подобных задач, постоянно совершенствуются, и мы внимательно следим за развитием подходов в области искусственного интеллекта, в том числе за исследованиями в обучении с подкреплением.

Обучение с подкреплением – это один из разделов машинного обучения, исследующий вычислительный подход к обучению агента, который пытается максимизировать свою совокупную накопленную награду путем взаимодействия со сложной, зачастую стохастической средой. Последние несколько лет исследования этого подхода переживают настоящий ренессанс – ни одна научная конференция по искусственному интеллекту не обходится без секции на эту тему. Каждый год публикуются сотни научных статей, и все больше компаний в России и за рубежом начинают применять последние достижения этой области в своем бизнесе для улучшения различных внутренних процессов – от рекомендательных систем до оптимизации цепей поставок.

Мы видим огромный потенциал практического применения методов обучения с подкреплением для совершенствования процессов в промышленности и логистике, а также верим в решающее значение данных теоретических концепций и алгоритмов для прогресса искусственного интеллекта как области человеческого знания. Несмотря на огромный интерес к этой области в последнее время, по указанной теме издано не так много литературы. Именно поэтому мы решили поучаствовать в публикации этой замечательной книги на русском языке.

Данная книга представляет собой исчерпывающее введение в такую интересную и быстро развивающуюся область искусственного интеллекта, как обучение с подкреплением. Ее авторы, Ричард Саттон и Эндрю Барто, проделали невероятную работу, описав простым и понятным языком не только ключевые концепции и алгоритмы обучения с подкреплением, но и современные достижения этой области. В книге продемонстрирована связь дисциплины с психологией и нейронауками. Авторами подробно рассматриваются детали работы системы AlphaGo, обыгравшей чемпиона мира в японскую настольную игру го, а также алгоритма, играющего в игры Atari на уровне человека, и многие другие приложения.

Мы желаем читателю удачи на пути изучения такой сложной, но невероятно полезной и увлекательной дисциплины.

Сергей Свиридов,
директор по исследованиям и разработкам, группа компаний «Цифра»

цифра

Группа компаний «Цифра» разрабатывает технологии цифровизации промышленности, инвестирует в продукты и развивает среду промышленного интернета вещей и искусственного интеллекта. Компания создала самую крупную в России лабораторию промышленного AI. Сегодня решения «Цифры» повышают эффективность промышленных предприятий в 22 странах мира. Ключевые отрасли для группы – это горная добыча и металлургия, машиностроение, нефтегазовый сектор и химическая промышленность. «Цифра входит» в Industrial Internet Consortium и ряд других российских и международных отраслевых ассоциаций.

Предисловие ко второму изданию

За двадцать лет, прошедших после выхода первого издания этой книги, мы стали свидетелями колоссального прогресса в области искусственного интеллекта, в немалой степени обусловленного достижениями машинного обучения, в т. ч. обучения с подкреплением. И этот прогресс был достигнут не только за счет впечатляющего роста вычислительных мощностей, но и благодаря развитию теории и алгоритмов. Поэтому необходимость во втором издании книги, вышедшей в 1998 году, давно назрела и перезрела, и наконец-то в 2012 году мы решили приняться за нее. Во втором издании мы ставили себе ту же цель, что и в первом: дать простое и понятное изложение основных идей и алгоритмов обучения с подкреплением, которое было бы доступно специалистам из смежных дисциплин. Книга по-прежнему осталась введением, основное внимание уделяется базовым алгоритмам онлайнового обучения. Мы включили ряд новых вопросов, возникших и приобретших важность за прошедшие годы, а также расширили описание тем, которые теперь понимаем лучше. Но мы даже не пытались дать исчерпывающее изложение всего предмета, который стремительно развивался во многих направлениях. Приносим извинения за то, что были вынуждены оставить все эти достижения (за исключением небольшого числа) без внимания.

Как и в первом издании, мы решили отказаться от строго формального изложения теории обучения с подкреплением и от постановки задачи в самом общем виде. Но по мере углубления нашего понимания некоторых вопросов потребовалось включить больше математики; части, для которых необходимо более уверенное владение математическим аппаратом, оформлены в виде врезок; читатели, не склонные к математике, могут их пропустить. Мы также используем не совсем такую же нотацию, как в первом издании. В процессе преподавания мы поняли, что новая нотация помогает устранить ряд распространенных недоразумений. Она подчеркивает различие между случайными величинами, которые обозначаются заглавными буквами, и их экземплярами, обозначаемыми строчными буквами. Например, состояние, действие и вознаграждение на временном шаге t обозначаются S_t, A_t и R_t, а их возможные значения – s, a и r. Кроме того, строчными буквами записываются функции ценности (например, v_π), а заглавными – их табличные представления (например, $Q_t(s, a)$). Приближенные функции ценности являются детерминированными функциями случайных параметров, поэтому также записываются строчными буквами (например, $\hat{v}(s, \mathbf{w}_t) \approx v_\pi(s)$). Векторы, например вектор весов $\mathbf{w}_t$ (ранее обозначался θ_t) и вектор признаков $\mathbf{x}_t$ (ранее ϕ_t), записываются строчными полужирными буквами, даже если являются случайными величинами. Заглавные полужирные буквы оставлены для матриц. В первом издании мы употребляли специальные обозначения $\mathcal{P}^a_{ss'}$ и $\mathcal{R}^a_{ss'}$ для вероятности перехода и ожидаемого вознаграждения. Один из недостатков этой нотации заключается в том, что она не полностью характеризует динамику вознаграждения,

а дает только математические ожидания – этого достаточно для динамического программирования, но не для обучения с подкреплением. Другой недостаток – чрезмерное количество верхних и нижних индексов. В этом издании мы ввели явное обозначение $p(s', r|s, a)$ для совместной вероятности следующего состояния и вознаграждения при условии текущего состояния и действия. Все изменения нотации сведены в таблице на стр. 20.

Второе издание значительно дополнено, и организация материала претерпела изменения. После первой вводной главы появились три новые части. В первой части (главы 2–8) обучение с подкреплением рассматривается настолько полно, насколько возможно без выхода за пределы табличного случая, для которого можно найти точные решения. Мы включили методы обучения и планирования для табличного случая, а также их унификацию в n-шаговых методах и в архитектуре Dyna. Многих алгоритмов, представленных в этой части, в первом издании не было, например: UCB, Expected Sarsa, двойное обучение, обновление по дереву, $Q(\sigma)$, RTDP и MCTS. Подробное рассмотрение табличного случая в начале книги позволяет изложить основные идеи в простейшей постановке. Вторая часть книги (главы 9–13) посвящена обобщению этих идей на аппроксимации функций. В ней появились новые разделы об искусственных нейронных сетях, о базисе Фурье, LSTD, ядерных методах, методах Gradient-TD и Emphatic-TD, методах среднего вознаграждения, истинно онлайновом методе TD(λ) и методах градиента стратегии. Во втором издании намного подробнее рассмотрено обучение с разделенной стратегией, сначала в табличном случае (главы 5–7), а затем для аппроксимации функций в главах 11 и 12. Еще одно отличие второго издания заключается в отделении идеи прямого представления, связанной с n-шаговым бутстрэппингом (теперь она более полно рассмотрена в главе 7), от идеи обратного представления, связанной со следами приемлемости (она теперь независимо описана в главе 12). В третью часть книги включены новые большие главы о связях обучения с подкреплением с психологией (глава 14) и нейронауками (глава 15), а также переработанная глава с примерами, включающая игры Atari, стратегию ставок в программе Watson, а также две программы игры в го: AlphaGo и AlphaGo Zero (глава 16). Но по необходимости мы смогли включить лишь малую часть сделанного в этой области. Выбор отражает наш давний интерес к недорогим безмодельным методам, которые хорошо масштабируются на крупные приложения. Последняя глава посвящена обсуждению будущего влияния обучения с подкреплением на общество. Хорошо это или плохо, но второе издание получилось почти в два раза больше первого.

Эта книга задумывалась как основной учебник для одно- или двухсеместрового курса по обучению с подкреплением. В односеместровый курс следует включить первые десять глав и излагать их по порядку. Это составит хорошую основу, к которой можно добавить материал из других глав, а также из других книг, например Bertsekas and Tsitsiklis (1996), Wiering and van Otterlo (2012), Szepesvári (2010), или из литературы – сообразуясь со вкусами лектора. В зависимости от подготовки студентов может оказаться полезным дополнительный материал по онлайновому обучению с учителем. Естественным дополнением будут идеи опций и моделей опций (Sutton, Precup and Singh, 1999). В двухсеместровый курс можно включить все главы и дополнительные материалы. Эту книгу можно также включить как часть более широких курсов машинного обучения, искусственного интеллекта или ней-

ронных сетей. В таком случае имеет смысл рассматривать только некоторое подмножество глав. Мы рекомендуем главу 1 в качестве краткого обзора, главу 2 до раздела 2.4, главу 3, а затем избранные разделы остальных глав в зависимости от располагаемого времени и интересов лектора и аудитории. Глава 6 наиболее важна для предмета и всей книги. В курс, ориентированный на машинное обучение или нейронные сети, следует включить главы 9 и 10, а в курс, ориентированный на искусственный интеллект или планирование, – главу 8. Разделы и главы, которые мы считаем более трудными и не существенными для книги в целом, помечены звездочкой. Их можно опустить при первом чтении без ущерба для понимания последующего текста. Упражнения повышенной сложности также помечены звездочкой, они не существенны для усвоения основного материала главы.

Большинство глав заканчиваются разделом «Библиографические и исторические замечания», в которых мы перечисляем источники идей, изложенных в главе, приводим ссылки на литературу для дальнейшего чтения и на текущие исследовательские работы, а также даем историческую справку. Несмотря на все усилия сделать эти разделы полными и авторитетными, мы наверняка упустили какие-то важные работы предшественников. Приносим свои извинения и открыты для исправлений и дополнений, которые будут внесены в электронную версию книги.

Это издание, как и первое, посвящено памяти А. Гарри Клопфа. Именно Гарри познакомил нас друг с другом, и именно его идеи о мозге и искусственном интеллекте побудили нас отправиться в долгое путешествие по миру обучения с подкреплением. Гарри получил образование в области нейрофизиологии и очень интересовался машинным интеллектом, он работал старшим научным сотрудником в отделе авионики Управления научно-исследовательских работ ВВС США (AFOSR) при базе ВВС Райт-Паттерсон в штате Огайо. Он был недоволен тем, что процессам поиска равновесия, в т. ч. гомеостазу и методам классификации на основе исправления ошибок, придают чрезмерно большую важность при объяснении естественного интеллекта и закладывания фундамента машинного интеллекта. Он отмечал, что системы, пытающиеся что-то максимизировать (не важно, что именно), качественно отличаются от систем поиска равновесия, и доказывал, что именно в максимизирующих системах ключ к пониманию важных аспектов естественного интеллекта и построения искусственного. Гарри сыграл решающую роль в получении от AFOSR финансирования для проекта оценки научной ценности этих и родственных им идей. Этот проект был запущен в конце 1970-х годов в Массачусетском университете в Амхерсте (UMass Amherst), сначала под руководством Майкла Эрбиба (Michael Arbib), Уильяма Килмера (William Kilmer) и Нико Спинелли (Nico Spinelli), профессоров факультета компьютерных и информационных наук и членов-основателей университетского кибернетического центра нейронаучных систем, созданного с перспективой работы на стыке нейронаук и искусственного интеллекта. Барто, недавно получивший докторскую степень в Мичиганском университете, был принят в проект на должность младшего научного сотрудника. Тем временем Саттон, студент старшего курса, изучавший информатику и психологию в Стэнфорде, переписывался с Гарри на тему их общего интереса к роли временных характеристик возбудителя в классической теории обусловливания. Гарри убедил группу в UMass в том, что Саттон станет отличным приобретением для проекта. Так Саттон оказался аспирантом

в UMass и начал писать докторскую диссертацию под руководством Барто, который к тому времени занял должность доцента. Исследования обучения с подкреплением, описанные в этой книге, – закономерный итог проекта, начатого Гарри и питавшегося его идеями. Таким образом, Гарри свел нас, авторов книги, положив начало долгой и плодотворной совместной работе. Посвящая эту книгу Гарри, мы отдаем должное его существенному вкладу не только в дисциплину обучения с подкреплением, но и в наше сотрудничество. Мы также выражаем благодарность профессорам Эрбибу, Килмеру и Спинелли за предоставленную нам возможность начать разработку этих идей. Наконец, мы благодарны AFOSR за щедрую поддержку, которую управление оказывало на ранней стадии наших исследований, и Национальному научному фонду (NSF) за щедрое финансирование в течение ряда последующих лет.

Есть много людей, которым мы благодарны за их идеи и помощь в подготовке второго издания. Все, кого мы благодарили за помощь в первом издании, заслуживают нашей глубочайшей благодарности и за это издание тоже – оно бы просто не состоялось без их вклада в первое издание. К этому длинному перечню мы обязаны добавить многих, кто помогал готовить только второе издание. Студенты, которым мы много лет преподавали эту дисциплину, отметились самыми разными способами: находили ошибки, предлагали исправления и – не в последнюю очередь – испытывали затруднения, заставляя нас думать, как объяснить материал лучше. Мы выражаем особую благодарность Марте Стинструп (Martha Steenstrup), которая прочитала весь текст и поделилась подробными комментариями. Главы по психологии и нейронаукам не были бы написаны без помощи многочисленных специалистов в этих областях. Мы признательны Джону Муру (John Moore) за его многолетние терпеливые разъяснения теории и экспериментов по обучению животных и основ нейронауки, а также за внимательное прочтение нескольких черновых вариантов глав 14 и 15. Мы также благодарны Мэтту Ботвинику (Matt Botvinick), Натаниэлю Доу (Nathaniel Daw), Питеру Дайяну (Peter Dayan) и Йелю Ниву (Yael Niv) за проницательные замечания к черновикам этих глав, помощь в освоении огромного массива литературы и указание на наши многочисленные ошибки в ранних вариантах рукописи. Разумеется, все оставшиеся ошибки в этих главах (а их не может не быть) – целиком наша вина. Мы выражаем благодарность Филу Томасу (Phil Thomas), который помог сделать эти главы доступными неспециалистам в области психологии и нейронаук, и Питеру Стерлингу (Peter Sterling), помогавшему сделать объяснения более понятными. Спасибо также Джиму Хоуку (Jim Houk) за знакомство с вопросами обработки информации в подкорковых ядрах головного мозга и за привлечение нашего внимания к смежным разделам нейронауки. Хосе Мартинес (José Martinez), Терри Сейновски (Terry Sejnowski), Дэвид Силвер (David Silver), Джерри Тезауро (Gerry Tesauro), Георгиос Теочарус (Georgios Theocharous) и Фил Томас (Phil Thomas) любезно помогли нам разобраться в деталях их приложений обучения с подкреплением, чтобы мы могли включить их в главу с примерами. Они же поделились ценными комментариями к черновым вариантам соответствующих разделов. Отдельное спасибо Дэвиду Силверу, который помог нам лучше понять дерево поиска Монте-Карло и программу DeepMind для игры в го. Мы также благодарны Джорджу Конидарису (George Konidaris) за помощь при написании раздела о базисе Фурье. Эмилио Картони (Emilio Cartoni), Томас Седерборг (Thomas Cederborg), Стефан Дернбах

(Stefan Dernbach), Клеменс Розенбаум (Clemens Rosenbaum), Патрик Тэйлор (Patrick Taylor), Томас Колин (Thomas Colin) и Пьер-Люк Бэкон (Pierre-Luc Bacon) помогали нам различными способами, за что мы им очень благодарны.

Саттон также выражает благодарность сотрудникам лаборатории обучения с подкреплением и искусственного интеллекта в университете Альберты за вклад во второе издание. Отдельное спасибо Рупаму Махмуду (Rupam Mahmood) за ценный вклад в обсуждение методов Монте-Карло обучения с разделенной стратегией в главе 5, Хамиду Мэю (Hamid Maei) за помощь в становлении взгляда на обучение с разделенной стратегией, представленного в главе 11, Эрику Грейвсу (Eric Graves) за постановку экспериментов в главе 13, Шан-тон Чжану (Shangtong Zhang) за воспроизведение и, как следствие, проверку почти всех экспериментальных результатов, Крису де Асису (Kris De Asis) за улучшение нового технического наполнения глав 7–12 и Харму ван Сейну (Harm van Seijen) за идеи, которые привели к отделению n-шаговых методов от следов приемлемости и (совместно с Хадо ван Хасселтом [Hado van Hasselt]) – за идеи, касающиеся точной эквивалентности прямого и обратного представления следов приемлемости (глава 12). Саттон также выражает признательность за финансовую поддержку и свободу исследований, которые обеспечивали правительство провинции Альберты и Национальный совет научных и инженерных исследований Канады на протяжении всей работы над вторым изданием книги. В частности, он благодарен Рэнди Гебелю (Randy Goebel) за создание благоприятной среды для исследований в Альберте с прицелом на перспективу. Также он благодарен компании DeepMind за поддержку на протяжении последних шести месяцев работы над книгой.

Наконец, мы признательны многочисленным придирчивым читателям черновых вариантов второго издания, которые мы выкладывали в интернет. Они нашли немало пропущенных нами ошибок и указали места, где может возникнуть недопонимание.

Предисловие к первому изданию

То, что теперь называется обучением с подкреплением, впервые привлекло наше внимание в конце 1979 года. Мы оба работали в Массачусетском университете над одним из ранних проектов воскрешения идеи сетей с нейроноподобными адаптивными элементами, которая могла оказаться многообещающим подходом к искусственному адаптивному интеллекту. Проект был посвящен исследованию «гетеростатической теории адаптивных систем» и разрабатывался под руководством А. Гарри Клопфа. Работа Гарри была богатейшим источником идей, а нам было позволено критически изучить их и сравнить с долгой историей предшествующих исследований в области адаптивных систем. Нашей задачей стало расчленение этих идей на составные части в попытке понять их взаимосвязи и сравнительную важность. Это продолжается и по сей день, но в 1979 году мы впервые осознали, что самая простая идея, которую долго считали чем-то само собой разумеющимся, удостоилась на удивление скромного внимания с вычислительной точки зрения. Это была идея обучающейся системы, которая хочет чего-то достичь и для этого адаптирует свое поведение так, чтобы максимизировать специальный сигнал со стороны окружающей среды. Иначе говоря, идея «гедонистической» обучающейся системы, или, как мы сказали бы теперь, идея обучения с подкреплением.

Как и многие другие, мы полагали, что обучение с подкреплением было всесторонне исследовано еще на заре развития кибернетики и искусственного интеллекта. Но при ближайшем рассмотрении оказалось, что его изучали очень поверхностно. Хотя обучение с подкреплением, безусловно, стало побудительным мотивом для некоторых ранних компьютерных исследований обучения, большая часть занимавшихся этим ученых затем обратились к другим вещам: классификации образов, обучению с учителем или адаптивному управлению, а то и вовсе забросили исследования в области обучения. В результате специальным вопросам, связанным с тем, как обучиться получать что-нибудь от среды, было уделено сравнительно мало внимания. Оглядываясь назад, можно сказать, что интерес к этой идее стал важнейшим шагом, приведшим в движение всю эту ветвь исследований. Мало чего можно было бы достичь в плане вычислительного обучения с подкреплением, не осознав, что столь фундаментальная идея ранее не была досконально исследована.

С тех пор эта область науки прошла долгий путь, развивалась в нескольких направлениях и стала зрелой дисциплиной. Обучение с подкреплением постепенно стало одним из самых активных направлений исследований в машинном обучении, искусственном интеллекте и нейронных сетях. Было подведено солидное математическое основание и созданы впечатляющие приложения. Компьютерные исследования обучения с подкреплением превратились в обширную область, в которой трудятся сотни ученых по всему миру, занимающиеся такими разными дисциплинами, как психология, теория управления, искусственный интеллект

и нейронауки. Особенно важны результаты, устанавливающие и развивающие связи с теорией оптимального управления и динамическим программированием. В целом проблема обучения путем взаимодействия ради достижения поставленных целей еще далека от решения, но наше понимание стало значительно глубже. Мы теперь можем изучать отдельные направления, например обучение на основе временных различий, динамическое программирование и аппроксимацию функций, в контексте их вклада в решение общей проблемы.

Принимаясь за написание книги, мы ставили цель дать простое и ясное описание ключевых идей и алгоритмов обучения с подкреплением. Мы хотели, чтобы изложение было доступно читателям, работающим в смежных дисциплинах, но не могли охватить их все одинаково подробно. В основном мы ведем изложение с точки зрения искусственного интеллекта и технического конструирования. Рассмотрение связей с иными областями мы оставляем другим авторам или отложим до следующего раза. Мы также решили отказаться от строго формального изложения предмета. Мы не стремились к максимальному уровню математического абстрагирования и не пытались доказывать теоремы. Мы постарались выбрать такой уровень математических деталей, который указал бы читателям с математическим складом ума верное направление, но не отвлекал от простоты и потенциальной общности базовых идей.

В некотором смысле мы работали над этой книгой тридцать лет, так что нам есть кого благодарить. Прежде всего мы благодарим людей, лично помогавших нам в разработке идей, представленных в книге: Гарри Клопфа (Harri Klopf), который помог нам осознать, что обучение с подкреплением нуждается в новой жизни; Криса Уоткинса (Chris Watkins), Димитрия Бертсекаса (Dimitri Bertsekas), Джона Цициклиса (John Tsitsiklis) и Пода Вербоса (Paul Werbos), которые помогли нам понять ценность связей с динамическим программированием; Джона Мура (John Moore) и Джима Кехоу (Jim Kehoe) за идеи из теории обучения животных; Оливера Селфриджа (Oliver Selfridge) за подчеркивание широты и важности адаптации; и вообще наших коллег и студентов, помогавших самыми разными способами: Рона Уильямса (Ron Williams), Чарльза Андерсона (Charles Anderson), Сатиндера Сингха (Satinder Singh), Сридхара Махадевана (Sridhar Mahadevan), Стива Брадтке (Steve Bradtke), Боба Крайтса (Bob Crites), Питера Дайяна (Peter Dayan) и Лимона Бэрда (Leemon Baird). На наши воззрения на обучение с подкреплением оказали большое влияние беседы с Полом Коэном (Paul Cohen), Полом Утгоффом (Paul Utgoff), Мартой Стинструп (Martha Steenstrup), Джерри Тезауро (Gerry Tesauro), Майком Джорданом (Mike Jordan), Лесли Кэлблингом (Leslie Kaelbling), Эндрю Муром (Andrew Moore), Крисом Аткисоном (Chris Atkeson), Томом Митчеллом (Tom Mitchell), Нильсом Нильсоном (Nils Nilsson), Стюартом Расселом (Stuart Russell), Томом Диттерихом (Tom Dietterich), Томом Дином (Tom Dean) и Бобом Нарендра (Bob Narendra).

Мы благодарны Майклу Литтману, Джерри Тезауро, Майклу Крайтсу, Сатиндеру Сингху и Вэй Чжану (Wei Zhang) за наполнение конкретикой разделов 4.7, 15.1, 15.4, 15.5 и 15.6 соответственно. Мы благодарим Управление научно-исследовательских работ ВВС США, Национальный научный фонд и лаборатории GTE за длительную финансовую поддержку, нацеленную на перспективу. Мы также выражаем признательность многим людям, которые читали черновые варианты книги и делились ценными замечаниями: Тому Калту (Tom Kalt), Джону Цициклису,

Павлу Чихошу (Pawel Cichosz), Олле Гэллмо (Olle Gällmo), Чаку Андерсону (Chuck Anderson), Стюарту Расселу, Бену ван Рою (Ben Van Roy), Полу Стинструпу (Paul Steenstrup), Полу Коэну, Сридхару Махадевану, Джетте Рандлов (Jette Randlov), Брайану Шеппарду (Brian Sheppard), Томасу О'Коннелу (Thomas O'Connell), Ричарду Коггинсу (Richard Coggins), Кристине Версино (Cristina Versino), Джону Х. Хайетту (John H. Hiett), Андреасу Баделту (Andreas Badelt), Джею Понте (Jay Ponte), Джо Беку (Joe Beck), Юстусу Пиатеру (Justus Piater), Марту Стинструп, Сатиндеру Сингху, Томми Яаколла (Jaakkola), Димитрию Бертсекасу (Dimitri Bertsekas), Торбьёрну Экману (Torbjörn Ekman), Кристине Бьёркман (Christina Björkman), Якобу Карлстрёму (Jakob Carlström) и Олле Палмгрену (Olle Palmgren). Наконец, мы благодарим Гвин Митчелл (Gwyn Mitchell) за разнообразную помощь, а также Гарри Стэнтона (Harry Stanton) и Боба Приора (Bob Prior), которые опекали нас в издательстве MIT Press.

Обозначения

Заглавными буквами обозначаются случайные величины, строчными – значения случайных величин и скалярные функции. Вещественные векторы записываются строчными полужирными буквами (даже если они являются случайными величинами), матрицы – заглавными полужирными буквами.

$\doteq$	равенство, имеющее место по определению
$\approx$	приближенное равенство
$\propto$	пропорционально
$\Pr\{X = x\}$	вероятность, что случайная величина X принимает значение x
$X \sim p$	случайная величина X выбрана из распределения $p(x) \doteq \Pr\{X = x\}$
$\mathbb{E}[X]$	математическое ожидание случайной величины X, т. е. $\mathbb{E}[X] \doteq \sum_x p(x)x$
$\operatorname{argmax}_a f(a)$	значение a, в котором $f(a)$ достигает максимума
$\ln x$	натуральный логарифм x
e^x	основание натуральных логарифмов, число $e \approx 2.71828$, возведенное в степень x; $e^{\ln x} = x$
$\mathbb{R}$	множество вещественных чисел
$f\colon \mathcal{X} \to \mathcal{Y}$	функция f, отображающая элементы множества $\mathcal{X}$ в элементы множества $\mathcal{Y}$
$\leftarrow$	присваивание
$(a, b]$	интервал вещественной оси между a и b, включающий b, но не включающий a
ε	вероятность предпринять случайное действие в ε-жадной стратегии
α, β	параметры, определяющие размер шага
γ	коэффициент обесценивания
λ	коэффициент затухания для следов приемлемости
$\mathbb{1}_{predicate}$	индикаторная функция ($\mathbb{1}_{predicate} \doteq 1$, если предикат *predicate* равен true, в противном случае 0)

В задаче о многоруких бандитах:

k	количество действий (рук)
t	дискретный временной шаг или номер игры
$q_*(a)$	истинное значение (ожидаемое вознаграждение) действия a
$Q_t(a)$	оценка $q_*(a)$ в момент t
$N_t(a)$	сколько раз действие a выбиралось до момента t
$H_t(a)$	обученное предпочтение действию a в момент t
$\pi_t(a)$	вероятность выбора действия a в момент t
$\bar{R}_t$	оценка ожидаемого вознаграждения в момент t при условии π_t

В марковском процессе принятия решений:

s, s'	состояния
a	действие
r	вознаграждение
$\mathcal{S}$	множество всех незаключительных состояний
$\mathcal{S}^+$	множество всех состояний, включая заключительное
$\mathcal{A}(s)$	множество всех действий, допустимых в состоянии s
$\mathcal{R}$	множество всех возможных вознаграждений, конечное подмножество $\mathbb{R}$
$\subset$	подмножество (например, $\mathcal{R} \subset \mathbb{R}$)
$\in$	элемент множества, например $s \in \mathcal{S}$, $r \in \mathcal{R}$
$\lvert\mathcal{S}\rvert$	количество элементов во множестве $\mathcal{S}$ (мощность $\mathcal{S}$)
t	дискретный временной шаг
T, $T(t)$	конечный временной шаг эпизода или эпизод, включающий временной шаг t
A_t	действие в момент t
S_t	состояние в момент t, обычно стохастически зависящее от S_{t-1} и A_{t-1}
R_t	вознаграждение в момент t, обычно стохастически зависящее от S_{t-1} и A_{t-1}
π	стратегия (правило принятия решения)
$\pi(s)$	действие, предпринятое в состоянии s при *детерминированной* стратегии π
$\pi(a \mid s)$	вероятность предпринять действие a в состоянии s при *стохастической* стратегии π
G_t	доход, начиная с момента t
h	горизонт, на какое время можно заглянуть вперед в прямом представлении
$G_{t:t+n}$,$G_{t:h}$	доход за n шагов с $t+1$ до $t+n$ или до h (обесцененный и скорректированный)
$\bar{G}_{t:h}$	плоский доход (необесцененный и нескорректированный) за шаги от $t+1$ до h (раздел 5.8)
G_t^λ	λ-доход (раздел 12.1)
$G_{t:h}^\lambda$	усеченный скорректированный λ-доход (раздел 12.3)
$G_t^{\lambda s}$, $G_t^{\lambda a}$	λ-доход, скорректированный на оценки ценности состояния или действия (раздел 12.8)
$p(s', r \mid s, a)$	вероятность перехода в состояние s' с вознаграждением r из состояния s после действия a
$p(s' \mid s, a)$	вероятность перехода в состояние s' из состояния s после действия a
$r(s, a)$	ожидаемое немедленное вознаграждение в состоянии s после действия a

$r(s, a, s')$ ожидаемое немедленное вознаграждение при переходе из s в s' после действия a

$v_\pi(s)$ ценность состояния s при стратегии π (ожидаемый доход)

$v_*(s)$ ценность состояния s при оптимальной стратегии

$q_\pi(s, a)$ ценность выполнения действия a в состоянии s при стратегии π

$q_*(s, a)$ ценность выполнения действия a в состоянии s при оптимальной стратегии

V, V_t массив оценок функции ценности состояний v_π или v_*

Q, Q_t массив оценок функции ценности действий q_π или q_*

$\bar{V}_t(s)$ ожидаемая приближенная ценность действия, например $V_t(s) \doteq \sum_a \pi(a|s) Q_t(s,a)$

U_t цель для оценки в момент t

δ_t ошибка временного различия (TD-ошибка) в момент t (случайная величина) (раздел 6.1)

δ_t^s, δ_t^a формы TD-ошибки для состояния и действия (раздел 12.9)

n в n-шаговых методах n – количество шагов бутстрэппинга

$\|v\|_\mu^2$ μ-взвешенная квадратичная норма функции ценности, $\|v\|_\mu^2 \doteq \sum_{s \in \mathcal{S}} \mu(s) v(s)^2$

d размерность – количество элементов $\mathbf{w}$

d' альтернативная размерность – количество элементов θ

$\mathbf{w}$, $\mathbf{w}_t$ d-мерный вектор весов, определяющий приближенную функцию ценности

w_i, $w_{t,I}$ i-й элемент обучаемого вектора весов

$\hat{v}(s, \mathbf{w})$ приближенная ценность состояния s при условии вектора весов $\mathbf{w}$

$v_\mathbf{w}(s)$ альтернативное обозначение $\hat{v}(s, \mathbf{w})$

$\hat{q}(s, a, \mathbf{w})$ приближенная ценность пары состояние–действий s, a при заданном векторе весов $\mathbf{w}$

$\nabla \hat{v}(s, \mathbf{w})$ вектор-столбец частных производных $\hat{v}(s, \mathbf{w})$ по $\mathbf{w}$

$\nabla \hat{q}(s, a, \mathbf{w})$ вектор-столбец частных производных $\hat{q}(s, a, \mathbf{w})$ по $\mathbf{w}$

$\mathbf{x}(s)$ вектор признаков, видимых в состоянии s

$\mathbf{x}(s, a)$ вектор признаков, видимых, когда в состоянии s предпринимается действие a

$x_i(s)$, $x_i(s, a)$ i-й элемент вектора $\mathbf{x}(s)$ или $\mathbf{x}(s, a)$

$\mathbf{x}_t$ сокращенное обозначение $\mathbf{x}(S_t)$ или $\mathbf{x}(S_t, A_t)$

$\mathbf{w}^\top \mathbf{x}$ скалярное произведение векторов, $\mathbf{w}^\top \mathbf{x} \doteq \sum_i w_i x_i$; например, $\hat{v}(\mathrm{s}, \mathbf{w}) \doteq \mathbf{w}^\top \mathbf{x}(s)$

$\mathbf{v}, \mathbf{v}_t$	вторичный d-мерный вектор весов, используемый для обучения $\mathbf{w}$ (глава 11)
$\mathbf{z}_t$	d-мерный вектор следов приемлемости в момент t (глава 12)
θ, θ_t	вектор параметров целевой стратегии (глава 13)
$\pi(a\|s, \theta)$	вероятность выбора действия a в состоянии s при условии параметрического вектора θ
π_θ	стратегия, соответствующая параметрам θ
$\nabla\pi(a\|s, \theta)$	вектор-столбец частных производных $\pi(a\|s, \theta)$ по θ
$J(\theta)$	мера качества для стратегии π_θ
$\nabla J(\theta)$	вектор-столбец частных производных $J(\theta)$ по θ
$h(s, a, \theta)$	предпочтение выбору действия a в состоянии s, основанное на θ
$b(a\|s)$	поведенческая стратегия, применяемая для выбора действий в процессе обучения целевой стратегии π
$b(s)$	базовая функция b: $\mathcal{S} \mapsto \mathbb{R}$ для методов градиента стратегии
b	коэффициент ветвления для МППР или дерева поиска
$\rho_{t:h}$	коэффициент выборки по значимости для временных шагов от t до h (раздел 5.5)
ρ_t	коэффициент выборки по значимости для одного только шага t, $\rho_t = \rho_{t:t}$
$r(\pi)$	среднее вознаграждение (коэффициент вознаграждения) для стратегии π (раздел 10.3)
$\bar{R}_t$	оценка $r(\pi)$ в момент t
$\mu(s)$	распределение состояний с единой стратегией (раздел 9.2)
μ	$\|\mathcal{S}\|$-мерный вектор $\mu(s)$ для всех $s \in \mathcal{S}$
$\|v\|^2_\mu$	μ-взвешенная норма функции ценности v, т. е. $\|v\|^2_\mu \doteq \sum_s \mu(s)v(s)^2$ (раздел 11.4)
$\eta(s)$	ожидаемое количество посещений состояния s в одном эпизоде (стр. 240)
Π	оператор проекции для функций ценности (стр. 316)
B_π	оператор Беллмана для функций ценности (раздел 11.4)
$\mathbf{A}$	матрица $\mathbf{A} \doteq \mathbb{E}[\mathbf{x}_t(\mathbf{x}_t - \gamma\mathbf{x}_{t+1})^\intercal]$ размерности $d \times d$
$\mathbf{b}$	d-мерный вектор $\mathbf{b} \doteq \mathbb{E}[R_{t+1}\mathbf{x}_t]$
$\mathbf{w}_{\text{TD}}$	неподвижная точка TD $\mathbf{w}_{\text{TD}} \doteq \mathbf{A}^{-1}\mathbf{b}$ (d-мерный вектор, раздел 9.4)
$\mathbf{I}$	единичная матрица
$\mathbf{P}$	матрица $\|\mathcal{S}\| \times \|\mathcal{S}\|$ вероятностей перехода состояний при стратегии π
$\mathbf{D}$	диагональная матрица $\|\mathcal{S}\| \times \|\mathcal{S}\|$ со значениями μ на диагонали
$\mathbf{X}$	матрица $\|\mathcal{S}\| \times d$ с векторами-строками $\mathbf{x}(s)$
$\overline{\text{VE}}(\mathbf{w})$	среднеквадратическая ошибка $\overline{\text{VE}}(\mathbf{w}) \doteq \|v_\mathbf{w} - v_\pi\|^2_\mu$ (раздел 9.2)
$\bar{\delta}_\mathbf{w}(s)$	беллмановская ошибка (математическое ожидание TD-ошибки), когда состоянием s является $v_\mathbf{w}$ (раздел 11.4)

$\bar{\delta}_{\mathbf{w}}$, BE беллмановский вектор ошибок с элементами $\bar{\delta}_{\mathbf{w}}(s)$

$\overline{\text{BE}}(\mathbf{w})$ среднеквадратическая беллмановская ошибка $\overline{\text{BE}}(\mathbf{w}) \doteq \|\bar{\delta}_{\mathbf{w}}\|_{\mu}^2$

$\overline{\text{PBE}}(\mathbf{w})$ среднеквадратическая спроецированная беллмановская ошибка $\overline{\text{PBE}}(\mathbf{w}) \doteq \|\prod\bar{\delta}_{\mathbf{w}}\|_{\mu}^2$

$\overline{\text{TDE}}(\mathbf{w})$ среднеквадратическая ошибка временных различий $\overline{\text{TDE}}(\mathbf{w}) \doteq \mathbb{E}_b[\rho_t {\delta_t}^2]$ (раздел 11.5)

$\overline{\text{RE}}(\mathbf{w})$ ошибка среднеквадратического дохода (раздел 11.6)

От издательства

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв прямо на нашем сайте **www.dmkpress.com**, зайдя на страницу книги, и оставить комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу **dmkpress@gmail.com**, при этом напишите название книги в теме письма.

Если есть тема, в которой вы квалифицированы, и вы заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу **http://dmkpress.com/authors/publish_book/** или напишите в издательство по адресу **dmkpress@gmail.com**.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы удостовериться в качестве наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг – возможно, ошибку в тексте или в коде, – мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от расстройств и поможете нам улучшить последующие версии данной книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу **dmkpress@gmail.com**, и мы исправим это в следующих тиражах.

Список литературы

Список использованной литературы лежит на сайте **dmkpress@gmail.com** на странице книги.

Нарушение авторских прав

Пиратство в интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и The MIT Press очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в интернете с незаконно выполненной копией любой нашей книги, пожалуйста, сообщите нам адрес копии или веб-сайта, чтобы мы могли применить санкции.

Пожалуйста, свяжитесь с нами по адресу электронной почты **dmkpress@gmail.com** со ссылкой на подозрительные материалы.

Мы высоко ценим любую помощь по защите наших авторов, помогающую нам предоставлять вам качественные материалы.

Глава 1

Введение

Идея о том, что мы обучаемся, взаимодействуя с окружающей средой, – вероятно, первое, что приходит в голову, когда мы задумываемся о природе обучения. Когда младенец играет, размахивает ручками или оглядывается вокруг, у него нет определенного учителя, однако имеется прямая сенсорно-двигательная связь со средой. Использование этой связи дает разнообразную информацию о причинах и следствиях, о последовательности действий и о том, что делать, чтобы достичь цели. На протяжении всей нашей жизни такие взаимодействия, несомненно, являются основным источником знаний об окружающей среде и о нас самих. Учимся ли мы водить машину или поддерживать беседу, мы всегда точно знаем, как среда реагирует на наши действия, и стремимся повлиять на происходящее с помощью своего поведения. Обучение посредством взаимодействия – фундаментальная идея, лежащая в основе почти всех теорий обучения и интеллекта.

В этой книге мы исследуем вычислительный подход к обучению посредством взаимодействия. Вместо того чтобы строить теории о том, как обучаются люди и животные, мы будем в основном рассматривать идеализированные ситуации и оценивать эффективность различных методов обучения[1]. То есть займем позицию исследователя или конструктора искусственного интеллекта. Мы будем изучать конструкции машин, которые эффективно решают проблемы обучения, представляющие научный или экономический интерес, оценивать эти конструкции методами математического анализа или с помощью компьютерных экспериментов. Принятый нами подход называется *обучением с подкреплением*, он в гораздо большей степени ориентирован на целеустремленное обучение посредством взаимодействия, чем другие подходы к машинному обучению.

1.1. Обучение с подкреплением

Обучение с подкреплением – это обучение тому, что делать, т. е. как отобразить ситуации на действия, чтобы максимизировать численный сигнал – вознаграждение. Обучаемому не говорят, какие действия предпринимать, он должен сам понять, какие действия приносят максимальное вознаграждение, пробуя их. В наиболее интересных и трудных случаях действия могут влиять не только на непосредственное вознаграждение, но и на следующую ситуацию, а значит, на все последующие вознаграждения. Эти две характеристики – поиск методом проб

[1] Связи с психологией и нейронауками излагаются в главах 14 и 15.

и ошибок и отложенное вознаграждение – являются наиболее важными отличительными чертами обучения с подкреплением.

Обучение с подкреплением, как и многие вещи, названия которых оканчиваются на «ние», например машинное обучение или скалолазание, одновременно является задачей, классом методов решения, хорошо работающих для этой задачи, и областью знаний, в которой изучается сама задача и методы ее решения. Удобно использовать одно название для всех трех вещей, понимая при этом, что концептуально они различны. В частности, различие между задачей и методами ее решения очень важно в обучении с подкреплением, и пренебрежение этим различием является источником множества недоразумений.

Мы формализуем задачу обучения с подкреплением, применяя идеи из теории динамических систем, а точнее как задачу оптимального управления не полностью известным марковским процессом принятия решений. Детали формализации подождут до главы 3, но основную мысль можно сформулировать просто – требуется уловить наиболее важные аспекты реальной проблемы, стоящей перед обучающимся агентом, который взаимодействует во времени с окружающей средой для достижения некоторой цели. Обучающийся агент должен уметь в какой-то степени воспринимать состояние среды и предпринимать действия, изменяющие это состояние. У агента также должна быть цель или несколько целей, как-то связанных с состоянием окружающей среды. Марковские процессы принятия решений включают все три аспекта – восприятие, действие и цель – в простейшей возможной форме, не сводя, однако, ни один аспект к тривиальному. Любой метод, подходящий для решения таких задач, будет рассматриваться нами как метод обучения с подкреплением.

Обучение с подкреплением отличается от *обучения с учителем*, еще одного вида обучения, который изучается в большинстве современных работ по машинному обучению. В случае обучения с учителем имеется обучающий набор помеченных примеров, подготовленный квалифицированным внешним учителем. Каждый пример представляет собой описание ситуации и спецификацию – метку – правильного действия, которое система должна предпринять в этой ситуации. Часто метка определяет категорию, которой принадлежит ситуация. Цель такого обучения – добиться, чтобы система смогла экстраполировать, или обобщить, свою реакцию на ситуации, которые не были предъявлены в обучающем наборе. Это важный вид обучения, но, взятый сам по себе, он не подходит для обучения с помощью взаимодействия. В интерактивных задачах часто практически невозможно получить примеры желаемого поведения, которые правильно представляли бы все ситуации, в которых агенту предстоит действовать. На неизведанной территории – там, где от обучения как раз и ожидают плодов, – агент должен уметь действовать, исходя из своего опыта.

Обучение с подкреплением отличается и от так называемого *обучения без учителя*, которое обычно имеет целью обнаружение структуры, скрытой в наборе непомеченных данных. Кажется, что термины «обучение с учителем» и «обучение без учителя» исчерпывают все возможные парадигмы машинного обучения, однако это не так. Может возникнуть соблазнительная мысль о том, что обучение с подкреплением – разновидность обучения без учителя, поскольку отсутствуют примеры правильного поведения. Но в действительности цель обучения с подкреплением – максимизировать вознаграждение, а не выявить скрытую структу-

ру. Выявление структуры в опыте агента, конечно, может быть полезно, но само по себе не решает задачу максимизации вознаграждения, стоящую перед обучением с подкреплением. Поэтому мы считаем обучение с подкреплением третьей парадигмой машинного обучения наряду с обучением с учителем, без учителя и, возможно, еще какими-то парадигмами.

Один из вызовов, стоящих перед обучением с подкреплением, но отсутствующий в других видах обучения, – нахождение компромисса между исследованием и использованием. Чтобы получить большое вознаграждение, обучающийся с подкреплением агент должен предпочитать действия, которые были испробованы в прошлом и принесли вознаграждение. Но чтобы найти такие действия, он должен пробовать действия, которые раньше не выбирал. Агент должен *использовать* уже приобретенный опыт, чтобы получить вознаграждение, но должен продолжать *исследования*, чтобы выбирать более эффективные действия в будущем. Дилемма состоит в том, что одного лишь исследования или использования недостаточно для успешного решения задачи. Агент должен пробовать разные действия и неуклонно отдавать предпочтение тем, которые кажутся наилучшими. В стохастической задаче каждое действие необходимо испробовать много раз, чтобы получить надежную оценку ожидаемого вознаграждения. Проблема исследования-использования интенсивно изучалась математиками на протяжении многих десятилетий, но и по сей день остается нерешенной. Пока что просто заметим, что вопрос о нахождении баланса между исследованием и использованием вообще не возникает в обучении с учителем и без учителя, по крайней мере в чистых вариантах этих парадигм.

Еще одна важная черта обучения с подкреплением – явное рассмотрение целостной проблемы целеустремленного агента, взаимодействующего с неопределенной окружающей средой. Этим оно отличается от многих подходов, в которых рассматриваются подзадачи, не задумываясь об их месте в общей картине. Например, мы упомянули, что во многих работах по машинному обучению изучается обучение с учителем, но явно не ставится вопрос о конечной пользе приобретенных способностей. Другие ученые разрабатывали теории планирования с целями общего вида, но не рассматривали роль планирования в принятии решений в режиме реального времени и не задавались вопросом, откуда возьмутся прогностические модели, необходимые для планирования. Хотя эти подходы принесли много полезных результатов, их сосредоточенность на изолированных подзадачах является серьезным ограничением.

В обучении с подкреплением принят противоположный подход – все начинается с полного интерактивного агента, преследующего некоторую цель. У всех обучающихся с подкреплением агентов имеются явные цели, все они могут воспринимать аспекты окружающей среды и выбирать действия, оказывающие влияние на среду. Более того, обычно с самого начала предполагается, что агент должен действовать, невзирая на значительную неопределенность окружающей его среды. Если обучение с подкреплением включает планирование, то оно должно учитывать взаимное влияние планирования и выбора действий в реальном времени, а также ответить на вопрос о том, откуда поступают и как совершенствуются модели. Если обучение с подкреплением включает обучение с учителем, то тому есть конкретные причины, которые определяют, какие способности критичны, а какие нет. Чтобы добиться прогресса в исследованиях по обучению, необ-

ходимо выделить и изучить важные подзадачи, но эти подзадачи должны играть понятные роли в полных интерактивных целеустремленных агентах, даже если детали полного агента еще только предстоит определить.

Под полным интерактивным целеустремленным агентом мы не всегда понимаем нечто вроде целостного организма или робота. Это возможные примеры, но полный интерактивный целеустремленный агент может быть также компонентом более крупной системы, наделенной определенным поведением. В таком случае агент прямо взаимодействует с остальными частями объемлющей системы и косвенно с окружением этой системы. Простой пример – агент, который следит за уровнем заряда батареи робота и посылает команды архитектуре управления роботом. Окружение этого агента – остальные части робота и окружение робота. Чтобы в полной мере оценить общность идеи обучения с подкреплением, нужно не ограничиваться очевидными примерами агентов.

Один из самых захватывающих аспектов современного обучения с подкреплением – его содержательные и плодотворные связи с другими инженерными и научными дисциплинами. Обучение с подкреплением – часть наблюдающейся уже несколько десятилетий тенденции к более тесной интеграции между искусственным интеллектом и машинным обучением, с одной стороны, и статистикой, оптимизацией и другими разделами математики – с другой. Например, способность некоторых методов обучения с подкреплением обучаться с помощью параметризованных аппроксиматоров решает классическую проблему «проклятия размерности» в исследовании операций и теории управления. Еще более заметна сильная связь обучения с подкреплением с психологией и нейронауками, приносящая ощутимые выгоды обеим сторонам. Из всех видов машинного обучения именно обучение с подкреплением ближе всего к способам обучения людей и животных, и истоки многих ключевых алгоритмов обучения с подкреплением следует искать в биологических самообучающихся системах. Обучение с подкреплением также вернуло долг – как в виде психологической модели обучения животных, лучше соответствующей некоторым эмпирическим данным, так и в виде влиятельной модели частей системы вознаграждения мозга. В этой книге развиваются идеи в основном обучения с подкреплением, относящиеся к конструированию и искусственному интеллекту, а связи с психологией и нейронауками описаны в главах 14 и 15.

Наконец, обучение с подкреплением – часть более широкой тенденции возвращения искусственного интеллекта к простым общим принципам. Начиная с конца 1960-х годов многие исследователи искусственного интеллекта предполагали, что нет никаких общих принципов, а интеллект обязан своим появлением разнообразным специальным приемам, процедурам и эвристикам. Иногда высказывалось мнение, что если бы мы могли заложить в машину достаточно фактов, скажем миллион или миллиард, то она обрела бы интеллект. Методы, основанные на общих принципах, таких как поиск или обучение, объявлялись «слабыми методами», тогда как основанные на специальных знаниях – «сильными методами». Эта точка зрения распространена и сегодня, но не является доминирующей. На наш взгляд, она просто была преждевременной: слишком мало усилий было вложено в поиск общих принципов, чтобы делать вывод об их отсутствии. В современных работах по искусственному интеллекту много внимания уделяется общим принципам обучения, поиска и принятия решений. Не ясно, насколько далеко назад

качнулся маятник, но обучение с подкреплением – безусловно, часть этого возвратного движения к простым и немногочисленным общим принципам искусственного интеллекта.

1.2. Примеры

Чтобы понять суть обучения с подкреплением, полезно рассмотреть несколько примеров и возможных приложений, которые стали стимулами для его разработки.

- Мастер-шахматист делает ход. Выбор продиктован планированием – предвидением возможных ответов и продолжений – и непосредственными интуитивными оценками желательности конкретных позиций и ходов.
- Адаптивный контроллер в режиме реального времени подстраивает параметры работы нефтеперерабатывающего завода. Контроллер оптимизирует соотношение между выходом готовой продукции, стоимостью и качеством, основываясь на заданных предельных затратах и не придерживаясь строго штатных режимов, предложенных конструкторами.
- Детеныш газели с трудом встает на ножки через несколько минут после рождения. Спустя полчаса он уже мчится со скоростью 20 миль в час.
- Подвижный робот решает, нужно ли ему войти в новое помещение для уборки мусора или уже пора искать станцию подзарядки. Он принимает решение, исходя из текущего уровня заряда батареи и того, насколько легко и быстро ему удавалось найти зарядное устройство в прошлом.
- Фил готовит завтрак. При ближайшем рассмотрении даже в этой, казалось бы, рутинной деятельности обнаруживается сложная паутина условных типов поведения и взаимных связей между целью и подцелями: подойти к кухонному шкафчику, открыть его, выбрать коробку с хлопьями, потянуться за ней, ухватить и вытащить. Другая не менее сложная, зависящая от внешних условий интерактивная последовательность действий необходима, чтобы достать тарелку, ложку и пачку молока. На каждом шаге мы наблюдаем многочисленные движения глаз для получения информации и управления передвижениями. Все время принимаются быстрые решения о том, как перемещать предметы и нужно ли поставить их на стол, перед тем как доставать другие. Каждый шаг диктуется целями, например взять ложку или открыть холодильник, и, в свою очередь, служит для достижения других целей, например иметь в распоряжении ложку, когда хлопья будут готовы, и в конечном итоге насытиться. Сознательно или нет, Фил получает информацию о состоянии собственного тела, на основании которой определяет, пора ли поесть, насколько он проголодался и что хочет покушать.

У этих примеров есть одна общая черта: они настолько привычны, что на них легко не обратить внимания. В каждом из них имеется *взаимодействие* между активным агентом, принимающим решения, и окружающей его средой, находясь внутри которой, агент стремится достичь *цели*, несмотря на *недетерминированность* среды. Действия агента могут влиять на будущее состояние среды (например, следующую позицию в шахматной партии, уровень заполнения хранилищ НПЗ, следующее местоположение робота и будущий уровень заряда его батареи) и тем самым на действия агента и доступные ему возможности в будущие момен-

ты времени. Для правильного выбора необходимо учитывать косвенные, отложенные последствия действий, поэтому может потребоваться прогнозирование или планирование.

При этом во всех рассмотренных примерах последствия действий невозможно предсказать достоверно; поэтому агент должен часто наблюдать за окружающей средой и действовать соответственно. Например, Фил должен следить, сколько молока он налил в тарелку с хлопьями, чтобы оно не пролилось на стол. Во всех примерах цели определены явно в том смысле, что агент может судить о продвижении к цели, полагаясь на свое непосредственное восприятие. Шахматист знает, выигрывает он или нет, контроллер НПЗ знает, сколько бензина произведено, детеныш газели знает, что упал, подвижный робот знает, когда заряд батареи подходит к концу, а Фил знает, завтракает он уже или нет.

Во всех приведенных примерах агент может использовать свой опыт для повышения результативности действий в будущем. Шахматист оттачивает интуицию при оценке позиций, а значит, улучшает свою игру; детеныш газели увеличивает эффективность бега; Фил учится не делать лишних движений во время приготовления завтрака. От знаний, с которыми агент приступает к выполнению задачи, – полученных из предыдущего опыта решения аналогичных задач или встроенных в него конструкторами или эволюцией, – зависит, чему полезно или легко обучиться, но только взаимодействие с окружающей средой позволит подкорректировать поведение для использования конкретных особенностей задачи.

1.3. Элементы обучения с подкреплением

Помимо агента и окружающей среды, можно выделить четыре главных элемента системы обучения с подкреплением: *стратегия*, *сигнал вознаграждения*, *функция ценности* и, факультативно, *модель* окружающей среды.

Стратегия определяет, как обучающийся агент поведет себе в данный момент времени. Грубо говоря, стратегия – это отображение множества воспринимаемых состояний среды на действия, предпринимаемые в этих состояниях. Она соответствует тому, что в психологии называется множеством правил, или ассоциаций, стимул–реакция. В некоторых случаях стратегия может быть простой функцией или таблицей соответствия, тогда как в других требует большого объема вычислений, например выполнения поиска. Стратегия лежит в основе обучающегося с подкреплением агента, поскольку ее одной достаточно для определения поведения. В общем случае стратегии могут быть стохастическими, т. е. задавать вероятности каждого действия.

Сигнал вознаграждения определяет цель в задаче обучения с подкреплением. На каждом временном шаге среда посылает обучающемуся агенту одно число, называемое *вознаграждением*. Единственное стремление агента – максимизировать полное вознаграждение, полученное в течение длительного времени работы. Таким образом, сигнал вознаграждения определяет, что для агента хорошо, а что плохо. В биологической системе аналогами вознаграждения являются удовольствие или боль. Вознаграждения – прямые и определяющие характеристики проблемы, стоящей перед агентом. Сигнал вознаграждения – главная причина изменения стратегии; если выбранное стратегией действие влечет низкое возна-

граждение, то стратегию следует изменить, так чтобы в будущем в такой ситуации выбиралось другое действие. В общем случае сигнал вознаграждения может быть стохастической функцией состояния среды и предпринятого действия.

Если сигнал вознаграждения показывает, что хорошо прямо сейчас, то *функция ценности* говорит, что хорошо в длительной перспективе. Грубо говоря, ценность состояния – это полное вознаграждение, которого агент может ожидать в будущем, если начнет работу в этом состоянии. Вознаграждение определяет непосредственную внутренне присущую желательность состояний окружающей среды, а ценность – долговременную желательность состояний с учетом тех состояний, которые с большой вероятностью встретятся позже, и вознаграждений в этих состояниях. Например, состояние может всегда приносить низкое немедленное вознаграждение, но при этом иметь высокую ценность, поскольку за ним регулярно следуют состояния, приносящие высокое вознаграждение. Обратное тоже может иметь место. Проводя аналогию с человеком, мы можем уподобить вознаграждение удовольствию (если оно высокое) или неудовольствию (если низкое), а ценность соответствует более продуманному и прозорливому суждению о том, насколько мы довольны или недовольны конкретным состоянием окружающей нас среды.

В некотором смысле вознаграждение первично, тогда как ценность, будучи прогнозом вознаграждения, вторична. Без вознаграждения не было бы ценности, а единственный смысл оценки ценности – получить большее вознаграждение. Тем не менее именно ценность стоит на первом месте, когда мы принимаем решения и оцениваем их последствия. Действия выбираются, исходя из суждений о ценности. Мы ищем те действия, которые приводят в состояния с наибольшей ценностью, а не те, которые приносят наибольшее вознаграждение, поскольку именно первые обещают максимальное вознаграждение в длительной перспективе. Вознаграждение мы получаем от среды непосредственно, а ценность необходимо оценивать и переоценивать, сообразуясь с последовательностями наблюдений, сделанных агентом за все время существования. На самом деле важнейшая часть почти всех рассматриваемых алгоритмов обучения с подкреплением – метод эффективного оценивания ценности. Центральная роль оценивания ценности – это, пожалуй, самое важное, что мы узнали об обучении с подкреплением за последние шестьдесят лет.

Четвертый и последний элемент некоторых систем обучения с подкреплением – *модель* окружающей среды. Это то, что имитирует поведение окружающей среды или, более общо, то, что позволяет делать выводы о том, как поведет себя среда. Например, зная состояние и действие, модель могла бы предсказать следующее состояние и следующее вознаграждение. Модели используются для *планирования*, под которым мы понимаем любой способ выбора порядка действий путем рассмотрения возможных будущих ситуаций, до того как они фактически произошли. Методы решения задач обучения с подкреплением, в которых используются модели и планирование, называются *основанными на модели*, в отличие от более простых *безмодельных* методов, в которых обучаемый явно действует методом проб и ошибок, считающимся чуть ли не полной *противоположностью* планированию. В главе 8 мы будем рассматривать системы обучения с подкреплением, которые одновременно обучаются методом проб и ошибок, в результате обучения строят модель окружающей среды и используют эту модель для планирования. Современное обучение с подкреплением охватывает весь спектр си-

стем – от низкоуровневого обучения методом проб и ошибок до высокоуровневого обоснованного планирования.

1.4. Ограничения и круг вопросов

Обучение с подкреплением в значительной мере опирается на понятие состояния – оно служит входной информацией для стратегии и функции ценности, а также входной и выходной информацией модели. Неформально можно считать состояние сигналом, доносящим до агента представление о том, как «выглядит окружающая среда» в конкретный момент времени. Формальное определение состояния дается в контексте марковских процессов принятия решения в главе 3. Но вообще мы призываем читателя опираться на неформальный смысл и рассматривать состояние как любую доступную агенту информацию об окружающей его среде. По существу, мы предполагаем, что сигнал состояния порождается какой-то системой предобработки, номинально являющейся частью окружающей агента среды. В этой книге мы не затрагиваем вопросы конструирования, изменения и обучения сигналу состояния (разве что очень кратко в разделе 17.3). Такой подход принят не потому, что мы считаем представление состояния чем-то несущественным, а чтобы полностью сосредоточиться на вопросах принятия решений. Иными словами, в этой книге нас интересует не конструирование сигнала состояния, а решение о том, какое действие предпринять при данном сигнале состояния.

Большинство рассматриваемых в книге методов обучения с подкреплением построено на оценивании функций ценности, но, вообще говоря, это необязательно. Например, в таких методах, как генетические алгоритмы, генетическое программирование, имитация отжига и другие методы оптимизации, функция ценности вообще не оценивается. В этих методах применяется несколько статических стратегий, каждая из которых на протяжении длительного времени взаимодействует со своим экземпляром среды. Стратегии, принесшие максимальное вознаграждение, и их случайные вариации, переходят в следующее поколение стратегий, после чего процесс повторяется. Мы называем такие методы *эволюционными*, потому что их работа напоминает то, как в результате биологической эволюции появляются организмы, обладающие осмысленным поведением, хотя они не обучались ему на протяжении собственной жизни. Если пространство стратегий достаточно мало или может быть организовано так, что хорошие стратегии встречаются часто или легко находятся (или если их поиску можно уделить много времени), то эволюционные методы могут оказаться эффективными. Кроме того, эволюционные методы имеют преимущества в задачах, где обучающийся агент не может воспринять полное состояние окружающей среды.

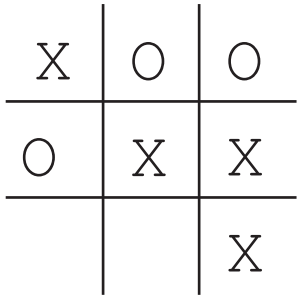
Наше внимание приковано к методам обучения с подкреплением, которые обучаются, взаимодействуя со средой, чего эволюционные методы не умеют. Методы, способные извлекать пользу из деталей поведения в отдельных актах взаимодействия, во многих случаях оказываются гораздо эффективнее эволюционных методов. Эволюционные методы игнорируют значительную часть полезной структуры, присутствующей в задаче обучения с подкреплением: они не используют тот факт, что искомая стратегия является функцией, отображающей состояния в действия; они не замечают, через какие состояния проходит индивидуум

на протяжении своей жизни и какие действия он выбирает. В некоторых случаях эта информация может сбивать с толку (например, когда состояния восприняты неправильно), но чаще повышают эффективность поиска. Хотя эволюция и обучение имеют много общего и естественно работают рука об руку, мы не считаем, что эволюционные методы сами по себе хорошо подходят для задач обучения с подкреплением, и потому не рассматриваем их в этой книге.

1.5. Развернутый пример: игра в крестики-нолики

Чтобы проиллюстрировать общую идею обучения с подкреплением и сравнить ее с другими подходами, мы подробно рассмотрим один пример.

Возьмем знакомую с детства игру в крестики-нолики. Два игрока по очереди делают ходы на доске 3×3. Один игрок ставит крестики (X), другой – нолики (O), до тех пор пока три одинаковых значка не выстроятся по горизонтали, по вертикали или по диагонали. Если доска заполнена, но трех значков подряд нет, считается, что игра закончилась вничью. Поскольку опытный игрок всегда может избежать проигрыша, предположим, что мы играем против неумехи, который иногда делает неправильные ходы и дает нам возможность выиграть. На некоторое время будем считать, что проигрыш и ничья одинаково плохи для нас. Как сконструировать игрока, который будет находить изъяны в игре противника и обучаться максимизировать свои шансы на выигрыш?



Это простая задача, но и ее трудно решить, применяя классические методы. Например, классическое «минимаксное» решение из теории игр не годится, потому что в нем предполагается, что противник играет определенным способом. Так, минимаксный игрок никогда не перейдет в состояние игры, в котором мог бы проиграть, даже если в действительности он всегда выигрывает в этом состоянии из-за неправильной игры противника. Классические методы оптимизации для задач с последовательным принятием решений, например динамическое программирование, могут вычислить оптимальное решение для любого противника, но требуют, чтобы на вход было подано полное описание противника, в частности вероятности выбора им каждого хода в любом состоянии на доске. Предположим, что эта информация заранее недоступна, именно так обстоит дело в большинстве задач, представляющих практический интерес. С другой стороны, ее можно оценить на основе опыта, сыграв много игр с противником. Едва ли не лучшее, что можно сделать в этой задаче, – сначала обучить модель поведения противника с некоторым уровнем доверия, а затем применить динамическое программирование для нахождения оптимального решения при данной приближенной модели противника. В конечном итоге это не так уж сильно отличается от некоторых методов обучения с подкреплением, которые мы будем изучать в данной книге.

Применение эволюционного метода к этой задаче означало бы, что мы должны произвести прямой поиск в пространстве возможных стратегий в попытке найти стратегию с высокой вероятностью выигрыша у противника. Здесь стратегией является правило, сообщающее игроку, какой ход сделать в каждом состоянии игры,

т. е. любой из возможных конфигураций крестиков и ноликов на доске 3×3. Для каждой рассматриваемой стратегии нужно было бы получить оценку вероятности выигрыша, сыграв несколько партий с противником. В результате такого вычисления мы получили бы указание, какую стратегию или стратегии рассматривать на следующем шаге. Типичный эволюционный метод осуществлял бы «восхождение на гору» в пространстве стратегий, последовательно генерируя и оценивая стратегии в попытке добиться постепенного улучшения. Или можно было бы использовать генетический алгоритм, который хранил бы и оценивал популяцию стратегий. В общем, можно было бы применить буквально сотни различных методов оптимизации.

А вот как к игре в крестики-нолики можно было бы применить метод, в котором используется функция ценности. Сначала нужно было бы подготовить таблицу чисел, по одному для каждого возможного состояния игры. Каждое число было бы равно последней оценке вероятности выиграть, начав с этого состояния. Эту оценку мы считаем ценностью состояния, а всю таблицу – обученной функцией ценности. Состояние A ценнее, или «лучше», состояния B, если текущая оценка вероятности выигрыша в A больше, чем в B. Предположим, что мы всегда играем крестиками, тогда для всех состояний с тремя X подряд вероятность выигрыша равна 1, поскольку мы уже выиграли. Аналогично для всех состояний с тремя O подряд, а также для заполненной доски вероятность выигрыша равна 0, поскольку в них мы выиграть не можем. Начальную ценность всех остальных состояний мы полагаем равной 0.5, т. е. мы думаем, что шанс выиграть составляет 50 %.

Затем мы играем много игр с противником. Для выбора хода мы рассматриваем состояния, которые получаются в результате каждого возможного хода (их столько, сколько пустых позиций на доске), и ищем их ценности в таблице. По большей части мы делаем ход жадно, т. е. выбираем такой ход, который ведет в состояние с наибольшей ценностью, т. е. с наибольшей оценкой вероятности выигрыша. Но иногда мы выбираем ход случайно. Такие ходы называются разведочными, поскольку приводят в состояния, которые мы иначе могли бы никогда не посетить. Последовательность сделанных и рассмотренных ходов за всю игру можно наглядно изобразить, как показано на рис. 1.1.

В процессе игры мы изменяем ценность состояний, в которых оказывались. Мы хотим, чтобы они более точно оценивали вероятность выигрыша. Для этого мы «переносим» ценность состояния после каждого жадного хода в состояние до этого хода, как показывают стрелки на рис. 1.1. Точнее, текущая ценность предыдущего состояния обновляется, так чтобы стать ближе к ценности следующего состояния. Это можно сделать, немного сдвинув ценность предыдущего состояния в направлении ценности следующего состояния. Если обозначить S_t состояние до жадного хода, а S_{t+1} – состояние после этого хода, то обновление оценки ценности S_t, обозначаемой $V(S_t)$, можно записать в виде:

$$V(S_t) \leftarrow V(S_t) + \alpha[V(S_{t+1}) - V(S_t)],$$

где α – небольшой положительный коэффициент, называемый параметром *размера шага*, который влияет на скорость обучения. Это правило обновления – пример метода обучения на основе *временных различий*. Он называется так, потому что изменения зависят от разности $V(S_{t+1}) - V(S_t)$ между оценками в соседние моменты времени.

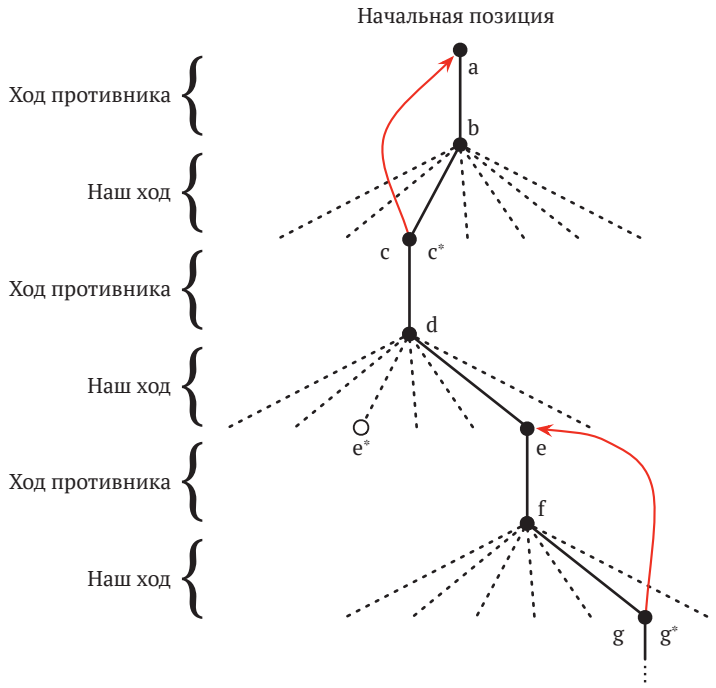


Рис. 1.1 ❖ Последовательность ходов в игре в крестики-нолики. Сплошными черными линиями показаны ходы, сделанные в игре; пунктирными линиями – ходы, которые мы (наш игрок, обучаемый с подкреплением) рассматривали, но не сделали. Второй ход был разведочным, т. е. он был сделан, несмотря на то что альтернативный ход (ведущий в состояние e*) оценивался выше. Разведочные ходы никак не сказываются на обучении, но все остальные ходы сказываются, а именно приводят к обновлениям, показанным красными линиями: оценки значений поднимаются вверх по дереву от более поздних узлов к более ранним, как объяснено в тексте

Описанный выше метод дает хорошие результаты на этой задаче. Так, если размер шага подходящим образом уменьшать со временем, то для любого фиксированного противника этот метод сходится к истинным вероятностям выигрыша в каждом состоянии при условии оптимальной игры нашим игроком. Более того, сделанные ходы (за исключением разведочных) действительно являются оптимальными ходами в игре против данного (неидеального) противника. Иными словами, метод сходится к оптимальной стратегии игры с данным противником. Если размер шага не уменьшается со временем до нуля, то этот игрок будет хорошо играть и с противниками, которые медленно изменяют манеру игры.

Этот пример иллюстрирует различие между эволюционными методами и методами, обучающими функцию ценности. Для вычисления стратегии эволюционный метод фиксирует стратегию и играет много игр с противником либо имитирует много игр с моделью противника. Частота выигрышей дает несмещенную оценку вероятности выиграть при данной стратегии и может использоваться при выборе следующей стратегии. Но любое изменение стратегии производится только после большого числа сыгранных игр, и используется лишь конечный результат каждой игры; что происходило *в процессе* игры, игнорируется. Например, если

игрок выиграл, то *все* его поведение в игре считается удачным независимо от того, насколько важными для выигрыша оказались отдельные ходы. Удачными считаются даже ходы, которые никогда не были сделаны! Напротив, методы на основе функции ценности позволяют оценивать отдельные состояния. В конечном итоге те и другие методы производят поиск в пространстве стратегий, но при обучении функции ценности задействуется информация, доступная по ходу игры.

Этот простой пример иллюстрирует некоторые ключевые особенности методов обучения с подкреплением. Во-первых, упор делается на обучение во взаимодействии с окружающей средой, в данном случае с противником по игре. Во-вторых, имеется ясная цель, и для выработки правильного поведения необходимо планирование или прогнозирование, учитывающее отложенные последствия выбора. Например, простой обучающийся с подкреплением игрок мог бы научиться расставлять многоходовые ловушки для близорукого противника. Поразительной особенностью обучения с подкреплением является тот факт, что оно может обеспечить эффект планирования и заглядывания в будущее, не используя модель противника и не выполняя явный поиск по всем возможным последовательностям будущих состояний и действий.

Хотя этот пример и иллюстрирует некоторые ключевые особенности обучения с подкреплением, он настолько прост, что может создаться впечатление, будто возможности обучения с подкреплением более ограничены, чем на самом деле. В крестики-нолики играют два человека, но обучение с подкреплением применимо и в случае, когда не существует внешнего противника, как, например, в «игре против природы». Кроме того, обучение с подкреплением не ограничивается задачами, в которых поведение распадается на отдельные эпизоды, как, например, серия игр в крестики-нолики, когда вознаграждение выплачивается только в конце каждого эпизода. В равной мере оно применимо и тогда, когда поведение не ограничено во времени, а вознаграждение разной величины можно получать в любой момент. Обучение с подкреплением применимо и к задачам, которые вообще не распадаются на дискретные временные шаги, как крестики-нолики. Общие принципы действуют и для непрерывных задач, хотя теория становится сложнее, и в этом введении мы ее опустим.

В крестиках-ноликах конечное и относительно небольшое множество состояний, а обучение с подкреплением можно использовать, и когда множество состояний очень велико и даже бесконечно. Например, в работах Gerry Tesauro (1992, 1995) описанный выше алгоритм объединен с искусственной нейронной сетью для игры в нарды, где количество состояний порядка 10^{20}. При таком числе состояний экспериментально исследовать можно только небольшую часть. Программа Тезауро научилась играть гораздо лучше всех предшествующих и в конце концов превзошла лучших в мире игроков-людей (см. раздел 16.1). Нейронная сеть наделяет программу способностью обобщать полученный опыт, так что в новых состояниях она выбирает ходы, опираясь на сохраненную информацию о похожих (с точки зрения сети) состояниях, наблюдавшихся в прошлом. Насколько хорошо система обучения с подкреплением может работать в задачах с таким обширным множеством состояний, сильно зависит от того, как успешно она обобщает предшествующий опыт. Именно для этого нам так нужно сочетание методов обучения с учителем и обучения с подкреплением. Искусственные нейронные сети и глубокое обучение (раздел 9.7) – не единственный и необязательно лучший способ добиться этого.

В примере с крестиками-ноликами обучение начиналось без какой-либо априорной информации, помимо знаний о правилах игры, но обучение с подкреплением ни коем случае не подразумевает взгляд на обучение и интеллект как на чистую доску. Напротив, априорную информацию можно включить в процесс обучения разными способами, от которых критически зависит эффективность обучения (см., например, разделы 9.5, 17.4 и 13.1). В крестиках-ноликах у нас был доступ к истинному состоянию, но обучение с подкреплением применимо и в случае, когда часть состояния скрыта или когда различные состояния воспринимаются обучаемым как одно.

Наконец, в крестиках-ноликах игрок может заглянуть вперед и узнать, какие состояния станут результатом каждого из возможных ходов. Для этого ему нужна модель игры, позволяющая предвидеть, как изменится среда в ответ на ходы, которые он, возможно, никогда не сделает. Таких задач много, но встречаются и другие, когда отсутствует даже краткосрочная модель последствий действий. Обучение с подкреплением применимо в обоих случаях. Модель необязательна, но если она имеется или может быть обучена, то ее легко использовать (глава 8).

С другой стороны, существуют методы обучения с подкреплением, которые вообще не нуждаются в модели окружающей среды. Безмодельные системы даже помыслить не могут о том, как изменится среда в ответ на одиночное действие. В этом смысле игрок в крестики-нолики является безмодельной системой по отношению к противнику: у него нет никакой модели противника. Поскольку модель полезна, только если достаточно верна, то безмодельные методы могут иметь преимущество над более сложными в случаях, когда трудность решения задачи связана именно со сложностью построения достаточно точной модели окружающей среды. Кроме того, безмодельные методы часто являются важными структурными элементами методов, основанных на модели. В этой книге мы посвятим несколько глав безмодельным методам и только потом обсудим, как использовать их в качестве компонентов более сложных методов на основе моделей.

Обучение с подкреплением можно использовать как на верхних, так и на нижних уровнях системы. В крестиках-ноликах игрок обучился лишь простейшим ходам, но ничто не мешает включить обучение с подкреплением в верхние уровни, где само «действие» заключается в применении некоторого, возможно, изощренного метода решения задач. В иерархических самообучающихся системах обучение с подкреплением может работать сразу на нескольких уровнях.

Упражнение 1.1. Игра с самим собой. Предположим, что вместо игры со случайным противником описанный выше алгоритм обучения с подкреплением играет против себя самого, и обе стороны обучаются. Как вы думаете, что произойдет в таком случае? Обучится ли игрок другой стратегии выбора ходов? □

Упражнение 1.2. Симметрии. Многие позиции в крестиках-ноликах выглядят различными, но на самом деле отличаются только симметрией. Как можно изменить описанный выше процесс обучения, чтобы воспользоваться этим фактом? Как это изменение могло бы улучшить процесс обучения? А теперь подумайте еще раз. Предположим, что противник не пользуется преимуществами симметрии. Должны ли тогда это делать мы? И верно ли, что позиции, совпадающие с точностью до симметрии, имеют одну и ту же ценность? □

Упражнение 1.3. Жадная игра. Предположим, что обучающийся с подкреплением игрок жадный, т. е. всегда выбирает ход, который переводит его в позицию с наивысшим рейтингом. Мог бы он научиться играть лучше нежадного игрока? Или хуже? Какие проблемы могли бы возникнуть? □

Упражнение 1.4. Обучение на результатах разведки. Предположим, что обучение подвергается обновлению после *любого* хода, включая разведочные. Если со временем размер шага уменьшается подходящим образом (но не уменьшается стремление к исследованию), то ценности состояний будут сходиться к другому набору вероятностей. Что концептуально представляют собой два набора вероятностей, вычисленных, когда мы обучаемся и не обучаемся на разведочных ходах? В предположении, что мы продолжаем совершать разведочные ходы, какой набор вероятностей предпочтительнее обучить? Какой принесет больше выигрышей? □

Упражнение 1.5. Другие улучшения. Можете ли вы предложить иные способы улучшить игрока, обучающегося с подкреплением? Можете ли вы придумать лучший способ решения задачи об игре в крестики-нолики в том виде, в котором она поставлена?

1.6. Резюме

Обучение с подкреплением – вычислительный подход к пониманию и автоматизации обучения и принятия решений, направляемых стремлением к достижению цели. Оно отличается от других вычислительных подходов упором на обучение агента в процессе прямого взаимодействия с окружающей средой, без посредничества учителя и без полной модели среды. По нашему мнению, обучение с подкреплением – первая дисциплина, в которой серьезно изучаются вычислительные проблемы, возникающие при обучении в процессе взаимодействия с окружающей средой ради достижения долгосрочных целей.

В обучении с подкреплением используется формализм марковских процессов принятия решений для описания взаимодействия обучающегося агента со средой в терминах состояний, действий и вознаграждений. Этот формализм задуман как простой способ представить существенные особенности проблемы искусственного интеллекта, к которым относятся восприятие причин и следствий, восприятие неопределенности и недетерминированности и существование явных целей.

Понятия ценности и функции ценности – ключ к большинству методов обучения с подкреплением, рассматриваемых в этой книге. Мы считаем, что функции ценности важны для эффективного поиска в пространстве стратегий. Применение функций ценности отличает методы обучения с подкреплением от эволюционных методов, которые производят поиск непосредственно в пространстве стратегий, руководствуясь оценкой стратегий в целом.

1.7. История ранних этапов обучения с подкреплением

В ранней истории обучения с подкреплением есть два основных направления, которые долго и плодотворно развивались независимо, прежде чем переплелись

в современном обучении с подкреплением. Одно направление связано с обучением методом проб и ошибок, оно уходит корнями в психологию обучения животных. Это направление можно проследить в самых ранних работах по искусственному интеллекту, и оно привело к возрождению обучения с подкреплением в начале 1980-х годов. Второе направление связано с задачей оптимального управления и ее решением с помощью функций ценности и динамического программирования. По большей части оно не соприкасается с обучением. Оба направления были практически независимы, но оказались в какой-то мере взаимосвязаны в третьем, не столь отчетливо выделяющемся направлении, связанном с методами на основе временных различий типа того, что мы применили в примере с крестиками и ноликами. Все три направления сошлись в конце 1980-х годов, когда сформировалась современная дисциплина обучения с подкреплением в том виде, в каком она представлена в этой книге.

Направление, ставящее во главу угла обучение методом проб и ошибок, – то, с чем мы лучше всего знакомы и о чем будет больше всего сказано в этом кратком историческом очерке. Но прежде все-таки посвятим несколько слов оптимальному управлению.

Термин «оптимальное управление» вошел в обиход в конце 1950-х годов и применялся для описания задачи о проектировании устройства управления, которое должно было минимизировать или максимизировать некоторую характеристику поведения динамической системы во времени. Один из подходов к решению этой задачи был разработан в середине 1950-х годов Ричардом Беллманом и другими учеными путем обобщения теории Гамильтона–Якоби, созданной в XIX веке. В этом подходе понятия состояния динамической системы и функции ценности, или «оптимальной функции выгоды», используются для вывода функционального уравнения, которое теперь часто называют уравнением Беллмана. Класс методов решения задач оптимального управления путем решения этого уравнения называется динамическим программированием (Bellman, 1957a). В работе Bellman (1957b) описана также дискретная стохастическая версия задачи оптимального управления, известная под названием «марковский процесс принятия решений» (МППР, англ. MDP). В работе Ronald Howard (1960) предложен метод итерации по стратегиям для МППР. Все это – существенные элементы, лежащие в основе теории и алгоритмов современного обучения с подкреплением.

Общепризнано, что динамическое программирование – единственный практически применимый способ решения общих стохастических задач оптимального управления. Оно страдает от того, что Беллман называл «проклятием размерности», т. е. требования к вычислительной мощности растут экспоненциально с ростом числа переменных состояния. Но все равно оно гораздо более эффективно и распространено, чем любой другой общий метод. Тема динамического программирования активно разрабатывалась с конца 1950-х годов, были предложены обобщения на частично наблюдаемые МППР (обзор см. в работе Lovejoy, 1991), многочисленные приложения (обзор см. в работах White, 1985, 1988, 1993), приближенные методы (обзор см. в работе Rust, 1996) и асинхронные методы (Bertsekas, 1982, 1983). Есть много отличных учебников по динамическому программированию (например, Bertsekas, 2005, 2012; Puterman, 1994; Ross, 1983; Whittle, 1982, 1983). В работе Bryson (1996) подробно изложена история оптимального управления.

Наличие связей между оптимальным управлением и динамическим программированием, с одной стороны, и обучением – с другой, осознавалось медленно. Мы не уверены, откуда взялось такое разделение, но, вероятно, основная причина – в разрыве между соответствующими дисциплинами и различии их целей. Возможно, свой вклад внесло преобладающее представление о динамическом программировании как пакетном методе вычислений, который сильно зависит от наличия точной модели системы и аналитических решений уравнения Беллмана. К тому же простейшая форма динамического программирования – вычисление, происходящее в обратном направлении по времени, поэтому трудно понять, как его можно применить в процессе обучения, который по необходимости протекает в прямом направлении. Про некоторые из самых первых работ по динамическому программированию, например Bellman and Dreyfus (1959), сегодня можно сказать, что они находятся в русле подхода к обучению. Работа Witten (1977), обсуждаемая ниже, точно может быть классифицирована как сочетание идей обучения и динамического программирования. В работе Werbos (1987) приводятся явные аргументы в пользу более тесной связи между динамическим программированием и методами обучения и доказывается, что динамическое программирование имеет прямое отношение к пониманию работы нейронов и когнитивных механизмов. Для нас полное слияние методов динамического программирования с онлайновым обучением случилось только после знакомства с работой Криса Уоткинса (Chris Watkins) 1989 года, в которой обучение с подкреплением излагалось с позиций формализма МППР. С тех пор эти связи активно разрабатывались многими исследователями, в особенности Димитрием Бертсеркасом и Джоном Цициклисом (Dimitri Bertsekas and John Tsitsiklis, 1996), которые предложили термин «нейродинамическое программирование», описывающий комбинацию динамического программирования с нейронными сетями. Еще один термин, широко употребляемый в настоящее время, – «приближенное динамическое программирование». В каждом из этих подходов на первое место выдвигаются различные грани предмета, но все они разделяют с обучением с подкреплением интерес к преодолению классических недостатков динамического программирования.

Мы считаем, что все работы по оптимальному управлению в какой-то мере являются работами по обучению с подкреплением. Мы определяем метод обучения с подкреплением как любой эффективный способ решения задач обучения с подкреплением, и теперь понятно, что эти задачи тесно связаны с задачами оптимального управления, особенно со стохастическими, подобными формулируемым в терминах МППР. Соответственно, мы должны считать методы решения задач оптимального управления, например динамическое программирование, также методами обучения с подкреплением. Поскольку почти все традиционные методы требуют полного знания об управляемой системе, кажется не вполне естественным говорить, что они часть обучения с подкреплением. С другой стороны, многие алгоритмы динамического программирования инкрементные и итеративные. Как и методы обучения, они постепенно приходят к правильному ответу путем последовательных приближений. Далее в этой книге мы покажем, что это отнюдь не поверхностное сходство. Теории и методы решения для случаев полного и неполного знания настолько тесно связаны, что, по нашему мнению, они должны рассматриваться вместе как части одного предмета.

А теперь вернемся к другому направлению, которое привело к современному обучению с подкреплением, – направлению, в основе которого лежит идея обучения методом проб и ошибок. Здесь мы затронем только основные положения, а более детальное рассмотрение отложим до раздела 14.3. Согласно американскому психологу Р. С. Вудворту (R. S. Woodworth, 1938), идея обучения методом проб и ошибок восходит к трудам Александра Бэна (Alexander Bain) 1850-х годов, в которых обсуждалось обучение «методом поисков и находок вслепую», а еще точнее – к британскому этологу и психологу Конвею Ллойду Моргану (Conway Lloyd Morgan), который в 1894 году употребил этот термин для описания своих наблюдений за поведением животных. Быть может, первым, кто кратко выразил существо обучения методом проб и ошибок, был Эдвард Торндайк (Edward Thorndike):

> Из нескольких возможных реакций на одну и ту же ситуацию, при прочих равных условиях, более тесно окажутся связанными с ситуацией те, что сопровождаются сразу или в недалеком будущем удовлетворенностью животного, поэтому, когда ситуация повторится, с большой вероятностью повторятся и эти реакции. А те реакции, которые сопровождаются сразу или в недалеком будущем неудовлетворенностью животного, будут иметь ослабленную связь с ситуацией, поэтому, когда ситуация повторится, эти реакции повторятся с меньшей вероятностью. Чем сильнее удовлетворенность или неудовлетворенность, тем сильнее или слабее связь (Thorndike, 1911, p. 244).

Торндайк называл это «законом эффекта», поскольку он описывает эффект, который подкрепляющие события оказывают на тенденцию выбора действий. Впоследствии Торндайк модифицировал этот закон, чтобы точнее учесть влияние последующих данных на обучение животного (например, различие между эффектами вознаграждения и наказания), и различные формы этого закона вызывали жаркие споры между теоретиками обучения (см., например, Gallistel, 2005; Herrnstein, 1970; Kimble, 1961, 1967; Mazur, 1994). Но, несмотря на это, закон эффекта – в той или иной форме – широко признан как основной принцип, определяющий многие виды поведения (Hilgard and Bower, 1975; Dennett, 1978; Campbell, 1960; Cziko, 1995). Это основа влиятельных теорий обучения Кларка Халла (Clark Hull, 1943, 1952) и общепринятых экспериментальных методов Б. Ф. Скиннера (B. F. Skinner, 1938).

Термин «подкрепление» в контексте обучения животных вошел в употребление спустя много лет после того, как Торндайк сформулировал закон эффекта, впервые он появился в этом контексте (насколько нам известно) в переводе на английский язык монографии Павлова об условных рефлексах, вышедшей в 1927 году. Павлов описывал подкрепление как усиление модели поведения вследствие получения животным стимула – подкрепителя – в непосредственной временной связи с другим стимулом или с реакцией. Некоторые психологи развили идею подкрепления, включив в нее не только усиление, но и ослабление поведения, а идею подкрепителя дополнили возможностью пропуска или прекращения стимулирования. Подкрепитель может считаться таковым, только если усиление или ослабление продолжается после его удаления; стимул, который просто привлекает внимание животного или инициирует его поведение, не порождая устойчивых изменений, не является подкрепителем.

Идея реализации обучения методом проб и ошибок в компьютере принадлежит к числу самых ранних идей о возможности искусственного интеллекта. В отчете 1948 года Алан Тьюринг описал конструкцию «системы удовольствие–неудовольствие», работавшей в духе закона эффекта:

> При достижении конфигурации, в которой действие не определено, производится случайный выбор недостающих данных, соответствующая запись в порядке эксперимента вносится в описание и применяется. Если имеет место стимул неудовольствия, то все экспериментальные записи отменяются, а если удовольствия, то все они становятся постоянными (Turing, 1948).

Было сконструировано много хитроумных электромеханических устройств, демонстрирующих обучение методом проб и ошибок. Возможно, самое первое построил Томас Росс (Thomas Ross, 1933), оно умело находить выход из простого лабиринта и запоминало путь с помощью установки переключателей. В 1951 году У. Грей Уолтер построил вариант «механической черепахи» (Walter, 1950), способной к простым формам обучения. В 1952-м Клод Шеннон продемонстрировал мышь по кличке Тесей, которая применяла метод проб и ошибок для поиска выхода из лабиринта, при этом сам лабиринт запоминал приведшие к успеху направления с помощью магнитов и реле, расположенных под полом (см. также Shannon, 1951). В работе J. A. Deutsch (1954) описана машина для решения лабиринтов, основанная на его теории поведения (Deutsch, 1953), которая в некоторых отношениях напоминала основанное на модели обучение с подкреплением (глава 8). Марвин Мински в своей докторской диссертации (Marvin Minsky, 1954) рассматривал вычислительные модели обучения с подкреплением и описал конструкцию аналоговой машины, состоящей из компонентов, которые он называл SNARC'ами (Stochastic Neural-Analog Reinforcement Calculators (стохастические нейроаналоговые калькуляторы подкрепления) и которые моделировали изменяемые синаптические связи в мозге (глава 15). На сайте **cyberneticzoo.com** имеется много информации об этих и многих других электромеханических обучающихся машинах.

Построение электромеханических самообучающихся машин уступило место программированию цифровых компьютеров для осуществления разнообразных типов обучения, в т. ч. методом проб и ошибок. В работе Farley and Clark (1954) описано цифровое имитационное моделирование машины с нейронной сетью, обучающейся методом проб и ошибок. Но вскоре их научные интересы сместились в сторону обобщения и распознавания образов, т. е. от обучения с подкреплением к обучению с учителем (Clark and Farley, 1955). Это положило начало путанице в части, касающейся связей между этими типами обучения. Многие ученые полагали, что занимаются обучением с подкреплением, тогда как на самом деле изучали обучение с учителем. Например, такие пионеры искусственного интеллекта, как Rosenblatt (1962) и Widrow and Hoff (1960), безусловно, вдохновлялись обучением с подкреплением – они оперировали терминами «вознаграждение» и «наказание», – но при этом изучали системы обучения с учителем, пригодные для распознавания образов и перцептивного обучения. Даже в наши дни некоторые исследователи и авторы учебников преуменьшают или затушевывают различия между этими типами обучения. Например, в некоторых учебниках по нейронным сетям использовался термин «метод проб и ошибок» для описания сетей,

которые обучаются на примерах. Эта путаница понятна, поскольку в таких сетях информация об ошибках используется для обновления весов связей. Но при этом упускается из виду существенная характеристика обучения методом проб и ошибок – выбор действия на основе оценочной обратной связи без знания о том, какое действие правильно.

Отчасти из-за этой путаницы исследования по настоящему обучению методом проб и ошибок стали редкостью в 1960-х и 1970-х годах, хотя есть и заметные исключения. В 1960-х годах термины «подкрепление» и «обучение с подкреплением» начали использоваться в технической литературе для описания инженерных применений обучения методом проб и ошибок (см., например, Waltz and Fu, 1965; Mendel, 1966; Fu, 1970; Mendel and McClaren, 1970). Особенно большое влияние оказала работа Мински «На пути к созданию искусственного интеллекта» (Minsky, 1961), где обсуждались некоторые вопросы, относящиеся к обучению методом проб и ошибок, в т. ч. прогнозирование, ожидание и то, что он называл *базовой задачей распределения поощрения для сложных систем обучения с подкреплением*: как распределить поощрение между многими решениями, которые могут быть причастны к успеху? Все обсуждаемые в этой книге методы в каком-то смысле направлены на решение этой задачи. Эта работа Мински заслуживает прочтения и сегодня.

В следующих абзацах мы обсудим ряд других исключений и частичных исключений из царившего в тот период пренебрежения к изучению истинного обучения методом проб и ошибок с вычислительной и теоретической точек зрения.

Одним из исключений стала работа новозеландского ученого Джона Андрэ (John Andreae), разработавшего систему STeLLA, которая обучалась методом проб и ошибок во взаимодействии с окружающей средой. Система включала внутреннюю модель мира, а впоследствии еще и «внутренний монолог» для решения проблемы скрытого состояния (Andreae, 1963, 1969a, b). В более поздней работе (Andreae, 1977) упор сделан больше на обучение с учителем, но все равно включена тема обучения методом проб и ошибок, причем порождение новых событий провозглашалось одной из целей системы. Примечательной особенностью этой работы стал «процесс подсоса», более полно разработанный в работе Andreae (1998). Это была реализация механизма распределения поощрения, похожая на операцию обновления, описанную нами выше. К сожалению, его пионерские исследования не получили широкой известности и не оказали существенного влияния на последующие изыскания в области обучения с подкреплением. Имеется написанный недавно реферат (Andreae, 2017a, b).

Более влиятельной оказалась работа Дональда Мичи (Donald Michie). В 1961 и 1963 годах он описал простую систему обучения игре в крестики-нолики методом проб и ошибок, которую назвал MENACE (Matchbox Educable Naughts and Crosses Engine). В ней каждой возможной игровой позиции соответствовал спичечный коробок (matchbox), содержащий ряд разноцветных бусин для каждого возможного в этой позиции хода. Извлечение случайной бусины из коробка, соответствующего текущей позиции, определяло ход MENACE. По завершении игры в коробки добавлялись или убирались использованные бусины, таким образом, ходы MENACE получали вознаграждение или наказание. В работе Michie and Chambers (1968) описана еще одна обучающаяся с подкреплением система игры в крестики-нолики, названная GLEE (Game Learning Expectimaxing Engine), и контроллер

обучения с подкреплением BOXES. Авторы применяли BOXES к задаче обучения, состоявшей в балансировании стержня, шарнирно закрепленного на движущейся тележке. В ней сигнал неудачи поступал, только когда стержень падал или тележка достигала конца дорожки. Эта задача ранее рассматривалась в работе Widrow and Smith (1964), в которой использовались методы обучения с учителем, в предположении, что инструкции поступают от учителя, который уже умеет балансировать стержень. Решение, предложенное Мичи и Чемберсом, – один из лучших ранних примеров задачи обучения с подкреплением в условиях неполного знания. Оно оказало влияние на многие более поздние работы по обучению с подкреплением, в т. ч. и наши собственные исследования (Barto, Sutton, and Anderson, 1983; Sutton, 1984). Мичи настойчиво подчеркивал роль метода проб и ошибок и обучения как существенных составных частей искусственного интеллекта (Michie, 1974).

В работе Widrow, Gupta, and Maitra (1973) был модифицирован алгоритм минимальной среднеквадратической ошибки (Least-Mean-Square – LMS), впервые изложенный в работе Widrow and Hoff (1960). Его целью было породить правило обучения с подкреплением, которое позволяло бы обучаться на сигналах об успехе и неудаче, а не на специально подготовленных обучающих примерах. Они называли эту форму обучения «избирательным бутстрэппингом» и описывали как «обучение с критиком», в отличие от «обучения с учителем». Они проанализировали это правило и показали, как алгоритм можно было обучить игре в блэкджек. Это был единичный экскурс Уидроу в область обучения с подкреплением, основные его работы, гораздо более влиятельные, относятся к обучению с учителем. Наше употребление термина «критик» заимствовано из статьи Widrow, Gupta, and Maitra. В работе Buchanan, Mitchell, Smith, and Johnson (1978) независимо использовался тот же термин в контексте машинного обучения (см. также Dietterich and Buchanan, 1984), но в их понимании критик – это экспертная система, способная на большее, чем просто оценка качества работы.

Исследования по *самообучающимся автоматам* оказали более прямое влияние на направление, связанное с методом проб и ошибок, и проложили путь к современным работам по обучению с подкреплением. Это методы решения неассоциативной чисто селекционной задачи обучения, известной под названием *k-рукий бандит* – по аналогии с игральным автоматом, или «одноруким бандитом», только рычагов не один, а k (см. главу 2). Самообучающиеся автоматы – это простые машины с небольшим объемом памяти, которые повышают вероятность получения вознаграждения в таких задачах. Их изучению положила начало работа русского математика и физика М. Л. Цетлина, выполненная совместно с коллегами в 1960-х годах (опубликована посмертно в Tsetlin, 1973). Затем эта тема активно разрабатывалась в интересах техники (см. Narendra and Thathachar, 1974, 1989). Изучались в том числе *стохастические самообучающиеся автоматы*, т. е. методы обновления вероятностей действий на основе сигналов вознаграждения. Алгоритм Alopex (Algorithm of pattern extraction), хотя и созданный не в традиции стохастических самообучающихся автоматов (Harth and Tzanakou, 1974), является стохастическим методом выявления корреляций между действиями и подкреплениями. Он оказал влияние на некоторые наши ранние исследования (Barto, Sutton, and Brouwer, 1981). Предвестником стохастических самообучающихся автоматов были более ранние работы по психологии, начиная с попыток Уильяма Эстеса (William Estes, 1950) создать статистическую теорию обучения, которая

затем получила развитие в работах других ученых (см., например, Bush and Mosteller, 1955; Sternberg, 1963).

Статистические теории обучения, разработанные в психологии, были подхвачены исследователями в области экономики, что заложило в этой области направление, посвященное обучению с подкреплением. Эта работа началась в 1973 году приложением теории обучения Буша и Мостеллера к ряду классических экономических моделей (Cross, 1973). Одной из целей было изучение искусственных агентов, которые действовали бы в большей степени как реальные люди, чем традиционные идеализированные экономические агенты (Arthur, 1991). Этот подход был распространен на изучение обучения с подкреплением в контексте теории игр. Обучение с подкреплением в экономике развивалось в значительной мере независимо от ранних работ по обучению с подкреплением в искусственном интеллекте, хотя теория игр остается темой, вызывающей интерес в обеих областях (но она выходит за рамки этой книги). В работе Camerer (2011) обсуждается традиция обучения с подкреплением в экономике, а в работе Nowé, Vrancx, and De Hauwere (2012) приведен обзор предмета с точки зрения многоагентных обобщений подхода, описываемого в этой книге. Подкрепление в контексте теории игр во многом отличается от того обучения с подкреплением, которое используется в таких игровых программах, как крестики-нолики, шашки и прочие развлечения. Обзор этого аспекта обучения с подкреплением и игр см., например, в работе Szita (2012).

В работе John Holland (1975) намечены контуры общей теории адаптивных систем, основанных на селекционных принципах. В его ранних работах метод проб и ошибок рассматривался в основном в неассоциативной форме, как в эволюционных методах и *k*-руких бандитах. В 1976 и затем более подробно в 1986 году он описал *системы классификации*, настоящие системы обучения с подкреплением, включающие ассоциацию и функции ценности. Ключевым компонентом систем классификации Холланда был «алгоритм пожарной цепочки» для распределения поощрения, тесно связанный с алгоритмом на основе временных различий, который использовался в нашем примере игры в крестики-нолики и обсуждается в главе 6. Еще одним важным компонентом был *генетический алгоритм* – эволюционный метод, роль которого заключалась в эволюционном порождении полезных представлений. Системы классификации всесторонне изучались многими учеными и стали крупной ветвью исследований по обучению с подкреплением (см. обзор Urbanowicz and Moore, 2009), но генетические алгоритмы – которые сами по себе, на наш взгляд, не являются системами обучения с подкреплением – удостоились гораздо большего внимания, как и другие подходы к эволюционным вычислениям (см., например, Fogel, Owens and Walsh, 1966, и Koza, 1992).

Человеком, который больше всех сделал для возрождения интереса к методу проб и ошибок в обучении с подкреплением в контексте искусственного интеллекта, стал Гарри Клопф (Harry Klopf, 1972, 1975, 1982). Клопф осознал, что существенные аспекты адаптивного поведения теряются, поскольку исследователи, занимающиеся обучением, чуть ли не поголовно увлечены одним лишь обучением с учителем. Согласно Клопфу, при этом утрачиваются гедонистические аспекты поведения, стремление добиться каких-то результатов от окружающей среды, управлять средой для приближения к желаемой цели прочь от нежелательной (раздел 15.9). Это основная идея обучения методом проб и ошибок. Идеи Клопфа оказали очень сильное влияние на авторов, поскольку в процессе их оценки (Barto and

Sutton, 1981a) мы пришли к пониманию различия между обучением с учителем и обучением с подкреплением и решили заняться последним. Значительная часть ранних работ, выполненных нами и нашими коллегами, была направлена на доказательство того, что обучение с учителем и обучение с подкреплением – действительно разные вещи (Barto, Sutton, and Brouwer, 1981; Barto and Sutton, 1981b; Barto and Anandan, 1985). Другие исследования показали, что обучение с подкреплением можно применить к решению важных задач обучения нейронных сетей, в частности к порождению алгоритмов обучения многослойных сетей (Barto, Anderson, and Sutton, 1982; Barto and Anderson, 1985; Barto, 1985, 1986; Barto and Jordan, 1987).

Теперь обратимся к третьему направлению в истории обучения с подкреплением – обучению на основе временных различий. Такие методы обучения отличаются тем, что основаны на различиях между оценками одной и той же величины (например, вероятности выигрыша в крестики-нолики), сделанными в соседние моменты времени. Это направление не такое обширное и не так выделяется, как два других, но оно сыграло особенно важную роль в этой области, отчасти потому, что методы на основе временных различий – нечто новое и существующее только в обучении с подкреплением.

Истоки обучения на основе временных различий следует искать в том числе в психологии обучения животных и, в частности, в понятии вторичных подкрепителей. Вторичным подкрепителем называется стимул, который подавался в паре с первичным подкрепителем, например пищей или болью, и в результате приобрел сходные свойства в плане подкрепления. Мински (Minsky, 1954), возможно, был первым, кто понял, что этот психологический принцип мог бы оказаться важным для систем обучения искусственного интеллекта. В работе Arthur Samuel (1959) впервые был предложен и реализован метод обучения, включавший идеи временных различий, это была часть его знаменитой программы игры в шашки (раздел 16.2).

Сэмюэл не ссылался на работу Мински и на возможные связи с обучением животных. Он, по-видимому, черпал вдохновение из предположения Клода Шеннона (Claude Shannon, 1950) о том, что компьютер можно запрограммировать для игры в шахматы с использованием функции оценки и что, возможно, его игру можно улучшить, если модифицировать эту функцию динамически. (Не исключено, что эти идеи Шеннона оказали также влияние на Беллмана, но у нас нет никаких подтверждений данной гипотезы.) Мински в своей работе 1961 года «На пути к...» подробно обсуждал работу Сэмюэла и предлагал связь с теориями вторичного подкрепления в обучении естественных и искусственных систем.

Как мы говорили, за десять лет, последовавших за исследованиями Мински и Сэмюэла, в области обучения методом проб и ошибок было выполнено мало вычислительных работ и уж совсем никаких по обучению на основе временных различий. В 1972 году Клопф соединил обучение методом проб и ошибок с важным компонентом обучения на основе временных различий. Клопфа интересовали принципы, которые масштабировались бы на обучение в больших системах, а потому он был заинтригован понятием локального подкрепления, посредством которого части системы обучения могли бы подкреплять друг друга. Он развил идею «обобщенного подкрепления», согласно которой каждый компонент (номинально: каждый нейрон) рассматривает все свои входы в терминах подкрепления: возбудители – как вознаграждение, а ингибиторы – как наказание. Это не

совсем та идея, которую мы теперь знаем под названием обучения на основе временных различий, и, оглядываясь назад, можно сказать, что она дальше от него, чем работа Сэмюэла. С другой стороны, Клопф связал эту идею с обучением методом проб и ошибок и сопоставил ее с большим объемом эмпирических данных по психологии обучения животных.

В работах Sutton (1978a, b, c) идеи Клопфа получили дальнейшее развитие, особенно в части связей с теориями обучения животных, в которых описывались правила обучения, основанные на изменениях предсказаний, сделанных в соседние моменты времени. Саттон вместе с Барто уточнили эти идеи и разработали психологическую модель классического обусловливания, положив в ее основу обучение на базе временных различий (Sutton and Barto, 1981a; Barto and Sutton, 1982). Затем последовали другие оказавшие заметное влияние модели того же вида (например, Klopf, 1988; Moore et al., 1986; Sutton and Barto, 1987, 1990). Некоторые нейронаучные модели, разработанные в то время, хорошо интерпретируются в терминах обучения на основе временных различий (Hawkins and Kandel, 1984; Byrne, Gingrich, and Baxter, 1990; Gelperin, Hop_eld, and Tank, 1985; Tesauro, 1986; Friston et al., 1994), хотя в большинстве случаев никаких исторических связей не прослеживается.

На наши ранние работы по обучению на основе временных различий оказали большое влияние теории обучения животных и работа Клопфа. Связи со статьей Мински «На пути к...» и с программами игры в шашки Сэмюэла были осознаны лишь впоследствии. Но к 1981 году мы прекрасно знали обо всех вышеупомянутых работах предшественников в контексте направлений, связанных с методом проб и ошибок и идеей временных различий. В то время мы разработали метод использования обучения на основе временных различий в сочетании с обучением методом проб и ошибок, получивший название *архитектура исполнитель–критик*, и применили его к задаче балансирования стержня Мичи и Чамберса (Barto, Sutton, and Anderson, 1983). Этот метод был подробно исследован в докторской диссертации Саттона (Sutton, 1984) и обобщен на нейронные сети с обратным распространением в докторской диссертации Андерсона (Anderson, 1986). Примерно в то же время Холланд (1986) явно включил идеи временных различий в свои системы классификации в форме алгоритма пожарной цепочки. Ключевой шаг был сделан в работе Sutton (1988), где обучение на основе временных различий было отделено от управления и рассмотрено как общий метод прогнозирования. В той же работе был сформулирован алгоритм TD(λ) и доказаны некоторые его свойства сходимости.

Когда в 1981 году работа над архитектурой исполнитель–критик подходила к концу, мы наткнулись на статью Ian Witten (1977, 1976a), которая, похоже, является самой ранней публикацией на тему правила обучения на основе временных различий. Автор предложил метод, который мы теперь называем табличным TD(0), для использования в составе адаптивного контроллера для МППР. Эта работа была предложена для публикации в журнале в 1974 году и вошла также в докторскую диссертацию Виттена, написанную в 1976 году. Работа Виттена шла по следам ранних экспериментов Андрэ с системой STeLLA и других систем обучения методом проб и ошибок. Таким образом, статья Виттена 1977 года охватывала оба основных направления исследований по обучению с подкреплением – метод проб и ошибок и оптимальное управление – и при этом внесла весомый ранний вклад в обучение на основе временных различий.

Направления, связанные с временными различиями и оптимальным управлением, полностью слились в 1989 году, когда Крис Уоткинс разработал Q-обучение. Эта работа знаменовала обобщение и объединение предшествующих исследований по всем трем направлениям обучения с подкреплением. Пал Вербос (Paul Werbos, 1987) внес вклад в это объединение тем, что начиная с 1977 года ратовал за конвергенцию обучения методом проб и ошибок и динамического программирования. К моменту выхода работы Уоткинса наблюдался колоссальный рост количества исследований по обучению с подкреплением, в основном в области машинного обучения, но также в контексте искусственных нейронных сетей и искусственного интеллекта в широком смысле. 1992 год ознаменовался замечательным успехом программы игры в нарды Джерри Тезауро, которая привлекла дополнительное внимание к этой тематике.

За время, прошедшее с момента выхода первого издания этой книги, получила развитие и доныне процветает подобласть нейронауки, изучающая связи между алгоритмами обучения с подкреплением и обучением с подкреплением в нервной системе. Причина тому – необыкновенное сходство между поведением алгоритмов на основе временных различий и активностью нейронов мозга, продуцирующих дофамин, на что указывали многие исследователи (Friston et al., 1994; Barto, 1995a; Houk, Adams, and Barto, 1995; Montague, Dayan, and Sejnowski, 1996; and Schultz, Dayan, and Montague, 1997). В главе 15 приведено введение в эту увлекательную область обучения с подкреплением. Важных достижений в недавней истории обучения с подкреплением слишком много, чтобы можно было упомянуть их все в этом кратком очерке, но многие отмечаются в конце тех глав, где о них идет речь.

Библиографические замечания

За дополнительными сведениями об обучении с подкреплением в целом отсылаем читателя к книгам Szepesvári (2010), Bertsekas and Tsitsiklis (1996), Kaelbling (1993a) и Sugiyama, Hachiya, and Morimura (2013). Из книг, рассматривающих предмет с позиций теории управления или исследования операций, отметим Si, Barto, Powell, and Wunsch (2004), Powell (2011), Lewis and Liu (2012) и Bertsekas (2012). В обзоре Cao (2009) обучение с подкреплением поставлено в один ряд с другими подходами к обучению и оптимизации стохастических динамических систем. Три специальных выпуска журнала «Machine Learning» посвящены обучению с подкреплением: Sutton (1992a), Kaelbling (1996) и Singh (2002). Упомянем полезные обзоры: Barto (1995b); Kaelbling, Littman, and Moore (1996) и Keerthi and Ravindran (1997). В томе под редакцией Weiring and van Otterlo (2012) имеется отличный обзор недавних достижений.

1.2 Пример с завтраком Фила заимствован из работы Agre (1988).

1.5 Метод на основе временных различий, использованный в примере игры в крестики-нолики, разрабатывается в главе 6.