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)