Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

на правах рукописи

Каледин Максим Львович

Разработка и анализ алгоритмов для задачи оптимального управления и обучения с подкреплением

Резюме диссертации
на соискание степени
Кандидата Компьютерных наук

Москва - 2023

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики» и Эколь Политекник (École Polytechnique, Institut Polytechnique de Paris), Париж.

Научные руководители:

Денис Витальевич Беломестный, к.физ.-мат.н., профессор, НИУ ВШЭ (Москва, Россия), Университет Дуйсбург-Эссен (University of Duisburg-Essen, Эссен, Германия)

Эрик Мулин, PhD, профессор Политехнического университета Парижа (École Polytechnique, Institut Polytechnique de Paris) (Париж, Франция).

Введение

Стохастическое оптимальное управление очень часто встречается на практике: от финансов [28, 64] до инженерных приложений [9]. Недавно подобные задачи приобрели новое значение и новые постановки в свете развивающейся области обучения с подкреплением (Reinforcement learning, RL), которая представляет собой в каком-то смысле пересечение оптимального управления, статистики и машинного обучения [58].

Подобные задачи можно приблизиетльно определить следующим образом. Пусть $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\geq 0})$ – это вероятностное пространство с фильтрацией $(\mathcal{F}_t)_{t\geq 0}$. Зададим также множество $\mathcal{U}$ измеримых случайных процессов $U : \mathbb{R}_{\geq 0} \times \Omega \to \mathbb{R}^n$, называемых *управлениями*, и множество *управляемых процессов*

$$\mathcal{X} = \left\{ X_t^U : U \in \mathcal{U} \right\},$$

где для каждого управления U каждый элемент $(X_t^U)_{t\geq 0}$ – это принимающий значения в $\mathbb{R}^d$ и $(\mathcal{F}_t)_{t\geq 0}$-согласованный случайный процесс. Мы также задаём функционал $J : \mathcal{X} \to \mathbb{R}$, называя его *функционалом выигрыша.*

Определение 1. *Задача поиска* $U_* \in \operatorname{Arg\,max}_{U\in\mathcal{U}} J(X^U)$ *называется задачей стохастического оптимального управленя.*

На практике (особенно в RL, см. [58]) в качестве составной части некоторых алгоритмов требуется оценить данное управление или решающее правило, приводя нас к задаче оценки.

Определение 2. *Задача оценки* $J(X^U)$ *для данного в некоторой форме управления* U *называется задачей оценки.*

Разумеется, в такой общей абстрактной постановке мы не можем судить о существовании решений или их свойствах. Однако этот вопрос проясняется, когда мы рассматриваем более специальные постановки. В диссертации рассмотрены две: задача оптимальной остановки для диффузионного процесса и Марковский процессс принятия решений (Markov Decision Problem, MDP).

Задача 1. (Оптимальная остановка диффузионного процесса, [64, 28]) Пусть $T > 0$ и процесс X_t задаётся стохастическим дифференциальнм уравнением в смысле Ито для $t \in [0, T)$

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{1}$$

С заданным начальным условием $X_0^U = x_0 \in \mathbb{R}^d$, где функции

$$b : [0, T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^d, \quad \sigma : [0, T) \times \mathbb{R}^d \times \mathrm{U} \to \mathbb{R}^{d\times n}$$

непрерывны и удовлетворяют условию Липшица по второму аргументу и условиям линейного роста с константой K:

$$\|b(t, x, u)\|_2 + \|\sigma(t, x, u)\|_2 \leq K(1 + \|x\|_2 + \|u\|_2),$$

где $\|\cdot\|_2$ означает подходящую Евклидову 2-норму. В подобных предположениях известно, что существует и единственно сильное решение дифференциального уравнения. Пусть $g_t : \mathbb{R} \to \mathbb{R}$ для всех $t \in [0, T]$ – некоторая функция, называемая обычно *функцией выплат.* Рассмотрим агента, наблюдающего за процессом, в момент

$t' \in [0, T]$ ему известны значения X_t для всех $t \leq t'$. Его цель – выбрать момент τ, когда он совершает какое-то действие (останавливает процесс, как говорят) и получает выплату $g_\tau(X_\tau)$. Формально, мы хотим выбрать время остановки τ, принимающее значения в $[0, T]$, из некоторого множества доступных времён остановки $\mathcal{T}$, максимизирующее дисконтированный ожидаемый выигрыш агента:

$$\tau_* = \arg\max_{\tau \in \mathcal{T}} \mathbb{E}\left[g_\tau(X_\tau)\right].$$

Наиболее популярные на практике методы происходят из идей алгоритмов Лонгштаффа-Шварца(LS)[44] и Тситсиклиса-Ван Роя [67]. Эти методы используют принцип динамического программирования и аппроксимируют условные матожидания с помощью регрессии методом наименьших квадратов, используя фиксированное множество базисных функций, на каждом шаге обратной индукции. Лонгштафф и Шварц продемонстрировали работу своего подхода на численных экспериментах, а в статьях [18] и [75] были приведены основные результаты о сходимости.

Задача 2. (Марковский процесс принятия решений, Markov Decision Process, MDP, [58]) Пусть заданы произвольные множества $\mathcal{S}, \mathcal{A}$, назовём их *пространством состояний* и *пространством действий* и снабдим их структурой измеримых пространств. Зададим однородную Марковскую цепь в дискретном времени S_t следующим образом. Пусть Π – множество стохастических решающих правил (называемых также *политиками*) $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$; иными словами, каждая политика принимает состояние $s \in \mathcal{S}$ и возвращает вероятностное распределение на пространстве действий, обозначаемое как $\pi(\cdot|s)$. Зададим *переходное ядро* $P(\cdot|s, a)$ как вероятностное распределение на пространстве состояний при условии данных текущих состояния s и действия a. Задаим $S_0 = s_0$ почти наверное и эволюцию S_t к S_{t+1} по следующей схеме:

$$A_t \sim \pi(\cdot|S_t),$$
$$S_{t+1} \sim P(\cdot|S_t, A_t).$$

Рассмотрим детерминированную функцию наград $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Естественной иллюстрацией MDP является агент в среде, описывающейся состояниями из $\mathcal{S}$; агент в момент t должен совершить действие A_t согласно его политике, после чего он получает награду $R(S_t, A_t)$, а среда изменяет своё состояние, как указано выше. Задачей оптимального управления для MDP состоит в том, чтобы максимизировать по политике ожидаемую сумму дисконтированных наград

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(S_t, A_t)\right],$$

где $\gamma \in (0, 1)$ играет роль дисконтирующего фактора и горизонт T может быть конечным (задача с конечным горизонтом) или бесконечным (задача с бесконечным горизонтом), или даже случайным (эпизодическая задача). MDP – фундаментальная модель в обучении с подкреплением (Reinforcement Learning,RL), которое сейчас быстро развивается и имеет многообещающие и существующие приложения в большом количестве областей, таких как: искусственный интеллект для игр [69, 8, 55], управление энергосистемами зелёной энергетики [41, 26], производство и роботы [2]. Вполне естественно RL даёт инженеру новые инструменты для построения управления в любых задачах автоматизации [27].

Оценка политики – важная часть model-free алгоритмов (т.е. не моделирующих среду), основанных итерации политики, и она, как правило, основывается на различных схемах стохастической аппроксимации, впервые предложенной в [51]. Стохастическая аппроксимация сама по себе стала хорошо изученной техникой [7, 39, 12], однако RL даёт ей новый смысл в констексте новых постановок. Среди прочих, линейные схемы стохастической аппроксимации наиболее популярны в обучении с подкреплением, так как они приводят к методам оценки политики линейными функциями; особенно важными в этом контексте являются методы temporal difference (TD) learning [57], для которых известны некоторые теоретические оценки для конечного времени (неасимптотические) в статьях [56, 40, 10, 20].

Цель работы

Цель нашего исследования состоит в получении результатов по задачам, описанным выше.

1. В Главе 1.1, мы приводим анализ вычислительной сложности алгоритма взвешенного Монте-Карло (Weighted Stochastic Mesh, WSM), похожего на метод Броуди и Глассермана из [13], для задач оптимальной остановки в дискретном и непрерывном времени. Мы сравниваем WSM с другими популярными методами с помощью новой метрики сложности, так как по классичекой метрике, принятой в стохастической вычислительной математике, все алгоритмы имеют неприемлемо большую сложность и их сложно между собой сравнивать.

2. В Главе 1.2 мы получаем неасимптотические результаты сходимости для схем линейной стохастической аппроксимации с двумя масштабами в предположении Марковского шума. Подобная постановка в точности совпадает с постановкой задачи в классических алгоритмах оценки политики в MDP: temporal difference learning (TD(0) of [57]) и gradient temporal difference algorithms (GTD[59],GTD2, а также скорректированный TDC [60]). Проблема существующих результатов в том, что они не берут в расчёт марковскую природу данных, крайне естественную в практическом контексте MDP, или их предположения слишком ограничивающи.

3. Наконец, в Главе 1.3 мы предлагаем новый метод снижения дисперсии в градиентных (Policy-gradient) методах оптимизации политики, основываясь на результатах об эмпирической дисперсии из [5]. Цель, прежде всего, в том, чтобы получить алгоритм, способный дать дополнительный выигрыш по сравнению с классическим методом Advantage Actor-Critic(A2C) для построения контрольных переменных[61] и, также, дать некоторые теоретические гарантии насчёт снижения дисперсии.

Главные результаты

1. Касательно первой цели, мы представляем анализ сложности алгоритма WSM, рассматривая также случай неизвестной переходной плотности $p(x|y)$, которую, тем не менее, можно аппроксимировать. Мы предложили новую метрику для сравнения алгоритмов оптимальной остановки, названную индексом ST

(*semitractability index*), и сравнили с её помощью WSM и несколько популярных в практическом сообществе алгоритмов: LS-алгоритм [44] и метод квантизации QTM [4].

2. Мы предоставляем обновлённые скорости сходимости для линейной стохастической аппроксимации с двумя масштабами для случаев мартингального и Марковского шума. Наш анализ позволяет выбирать достаточно общие размеры шагов для обновления, включая константный, кусочно константный, и убывающий, исследованные в [33, 19, 74, 22]. В отличие от предыдущих работ по этой теме [42, 19, 74], наши результаты получены без использования дополнительных проекций для устойчивости. Наконец, при некоторых дополнительных предположениях на шаг, мы вычисляем точное асимптотическое разложение ожидаемой квадратичной ошибки, чтобы показать точность полученных верхних оценок.

3. Мы предлагаем два новых policy-gradient метода (EV-методы), основанных на EV-критерии (Empirical Variance), и демонстрируем их работу в нескольких практических задачах, сравнивая их с A2C. Также мы предоставляем теоретические оценки дисперсии оценки градиента с помощью идей [5]; в RL это первые оценки дисперсии с большой вероятностью, полученные с помощью статистической теории обучения. Измерения дисперсии градиентных оценок на экспериментах приводят в каком-то смысле необычным выводам. Во-первых, EV-методы способны решать задачу снижения дисперсии часто значительно лучше A2C. Во-вторых, мы наблюдаем подтверждения гипотезы [68]: снижение дисперсии действительно может приводить к ускорению обучения, но это сильно зависит от конструкции среды. Мы приводим первое эмпирическое исследование EV-критерия для policy-gradient методов в классических образцовых задачах и первую имплементацию подхода в контексте PyTorch.

Авторский вклад. Часть анализа для дискретного времени, перевод результатов дискретного времени в случай непрерывного времени, имплементация алгоритма и численные эксперименты в первой статье выполнены Автором. В статье 2 автор провёл большую работу по подготовке литературного обзора и написанию доказательств в мартингальном случае, а также предоставил численные результаты и примеры. По последнему направлению Автор провёл все главные шаги доказательств вероятностных оценок, предположений, а также много участвовал в литературном обзоре, имплементации алгоритма и проведении численных экспериментов.

Апробация и Публикации

Основные публикации, выносимые на защиту

1. Denis Belomestny, Maxim Kaledin, and John Schoenmakers. Semitractability of optimal stopping problems via a weighted stochastic mesh algorithm. *Mathematical Finance*, 30(4):1591–1616, 2020

2. Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third*

Conference on Learning Theory, volume 125 of *Proceedings of Machine Learning Research*, pages 2144–2203. PMLR, 09–12 Jul 2020

Прочие публикации

1. Maxim Kaledin, Alexander Golubev, and Denis Belomestny. Variance reduction for policy-gradient methods via empirical variance minimization. *arXiv:2206.06827v2*, 2022

Выступления на конференциях и доклады на семинарах

1. Kaledin M. *Variance Reduction for Policy-Gradient Methods via Empirical Variance Minimization*, summer school "Learning and Optimization in Artificial Intelligence Models HSE, Saint-Petersburg, June 20-26 2022.

2. Kaledin M. *Theoretical Analysis and Variance Reduction in Reinforcement Learning Algorithms*, CMAP Doctoral Student Reports, CMAP Institut Polytechnique de Paris, Palaiseau, France, May 31 2021.

3. Kaledin M. *Variance Reduction for policy-gradient methods in Reinforcement Learning*, PhD Research seminar of Doctoral School of Computer Science, HSE, Moscow, Russia, December 21 2020 .

4. Kaledin M. *Variance Reduction for policy-gradient methods in Reinforcement Learning*, summer school "Modern methods of Information Theory, Optimization and Control Sirius, Sochi, Russia, August 2-23 2020.

5. Kaledin M. *Convergence of Linear Two-Timescale Stochastic Approximation*, Winter School "Math of Machine Learning" , Sirius, Sochi, Russia, February 20-23 2019.

6. Kaledin M. *Approximate Dynamic Programming for American Options*, poster session, "Data Science Summer School"(DS3), l'École Polytechnique, Paris, June 24-28th 2019.

7. Kaledin M. *Approximate Dynamic Programming with Approximation of Transition Density*, Winter School "New Frontiers in High-Dimensional Probability and Statistics 2" , HSE, Moscow, February 22-23 2019.

1 Основные результаты

1.1 ST-решаемость задачи оптимальной остановки с помощью алгоритма взвешенного Монте-Карло

Результаты этой главы опубликованы в статье [6].

1.1.1 Введение

Задача оптимальной остановки состоит в построении решающего правила, говорящего, когда нужно осуществить решение ("остановить" процесс). Будучи классической задачей в финансовой математике, она в точности получается при оценке различных опционов, производных финансовых инструментов, среди которых самые популярные – американские и европейские [28]. Рассмотрим две постановки.

1. (Задача в непрерывном времени) Рассмотрим множество возможностей для остановки $[0, T]$ и пусть $(X_t)_{t\in[0,T]}$ – это, как в Задаче 1, диффузионный процесс заданный уравнением (1) в смысле Ито. Задача в точности такая же с g_t в роли функции выплат для каждого $t \in [0, T]$ и $\mathcal{T}$ – множеством возможных времён остановок.

2. (Задача в дискретном времени) Дискретная версия задачи выше, задано конечное множество возможностей для остановки $\mathcal{L} = \{0, .., L\}$ с $L \in \mathbb{Z}_{>0}$ и $(Z_l)_{l\in\mathcal{L}}$ – марковская цепь в $\mathbb{R}^d$, полученная после дискретизации по времени. Нужно найти время остановки τ^*, дающее

$$\mathbb{E}\left[g_{\tau_*}(Z_{\tau_*}) \;\middle|\; Z_0\right] = \sup_{\tau\in\mathcal{T}} \mathbb{E}\left[g_\tau(Z_\tau) \;\middle|\; Z_0\right],$$

   где g_l – функции выплат $\mathbb{R}^d \to \mathbb{R}_{\geq 0}$ в моменты $l \in \mathcal{L}$ и $\mathcal{T}$ – множество возможных времён остановок, принимающих значения в $\mathcal{L}$. Для простоты и не ограничивая общности мы полагаем марковскую цепь $(Z_l)_{l\in\mathcal{L}}$ однородной с одношаговой переходной плотностью $p(y|x)$, то есть,

$$\mathbb{P}\left(Z_{k+1} \in dy \mid Z_k = x\right) = p(y|x)dy$$

   для всех $x, y \in \mathbb{R}^d$.

Несмотря на существующие результаты о сходимости, оказывается, что сравнить различные алгоритмы для задачи оптимальной остановки, основываясь только на скоростях сходимости, невозможно, поскольку алгоритмы очень сильно отличаются в вычислительном аспекте. Основные подходы к анализу численных алгоритмов можно найти в [49] и ссылках там. Основная задача, изучаемая в литературе, – это вычисление интегралов с помощью детерминированных и стохастических алгоритмов. Оптимальная остановка, на самом деле, это вычисление нескольких в определённом смысле вложенных интегралов, получающихся в результате принципа динамического программирования. Следовательно, существующие результаты не могут быть просто применены к задаче оптимальной остановке. В частности, для LS-алгоритма [75, Cor. 3.10] вычислительная сложность составляет

$$\mathcal{C}_L(\varepsilon, d) \sim \kappa_1 \frac{L 5^{(\kappa_2+L)(2+3d/\alpha)}}{\varepsilon^{2+3d/\alpha}}$$

с определёнными константами κ_1, κ_2. Если рассматривать задачу в непрерывном времени, то тогда, подбирая дискретизацию по времени, мы приходим к сложности, которая может расти быстрее, чем $\exp(\varepsilon^{-1/\beta})$ для некоторого $\beta > 0$. Похожие оценки верны для других алгоритмов, основанных на симуляции и регрессии, включая алгоритм Тситсиклиса и Ван Роя [67]. В [24] рассмотрена более общая регрессионная схема с похожими результатами. Главная проблема этих оценок сложности состоит

в том, что размерность процесса d входит в показатель степени ε, приводя к так называемому *проклятию размерности*, всё ещё присутствующему даже в таких схемах Монте-Карло. Существует, однако, ещё работы [29], где предложена новаторская Монте-Карло схема с независимой от d сложностью, но, к сожалению, сложность экспоненциально зависит от ε^{-1}.

Решаемость (tractability) – важное понятие в анализе численных алгоритмов один из способов определить её следующий. Численная проблема в d-мерном пространстве, например, вычисление интеграла $\int_{[0,1]^d} f(x)dx$, называется *решаемой*(tractable) [49], если есть алгоритм, решающий её со сложностью $\mathcal{C}(\varepsilon, d)$, удовлетворяющей

$$\lim_{d+\varepsilon^{-1}\to\infty} \frac{\ln \mathcal{C}(\varepsilon, d)}{d + \varepsilon^{-1}} = 0. \tag{2}$$

В случае оптимальной остановки такое определение не очень полезно: все регрессионные алгоритмы уже в случае дискретного времени оказываются нерешаемыми; действительно,

$$\limsup_{d+\varepsilon\to\infty} \frac{\ln \mathcal{C}(\varepsilon, d)}{d + \varepsilon^{-1}} = \infty$$

в силу экспоненциальной зависимости сложности от d (согласно результатам, известным в литературе). Следовательно, даже для оптимальной остановки в дискретном времени регрессионные методы не приводят к решаемости, согласно классическому определению. Например, согласно результатам [65], может быть показано, что ошибка аппроксимации функции ценности в этом случае приобретает форму

$$5^L \left(\sqrt{\frac{m^d}{N}} + e^{-\theta m} \right), \quad \theta > 0.$$

Это наблюдение также применимо к методу взвешенного Монте-Карло(Weighted Stochastic Mesh, WSM), предложенному Броуди и Глассерманом в [13], таким образом делая почти все алгоритмы нерешаемыми. Этот факт мотивирует разработку более гибкого подхода к метрике сложности, которая бы позволяла сравнивать алгоритмы для задач оптимальной остановки.

Оказывается, что мало известно про свойства сходимости WSM-алгоритма, за исключением некоторых первых результатов в дискретном случае [1]. Авторы, к сожалению, не предоставляют зависимость ошибки от размерности и количество возможностей остановки и их анализ основан на достаточно ограничивающем предположении компактного пространства состояний. Алгоритм, похожий на тот, что мы анализируем, был также представлен Растом в [52], где представлена Монте-Карло схема без экспоненциальной зависимости от d, но лишь $O(1/\varepsilon^4)$. Постановка с MDP в дискретном времени и использованные техники делают нетривиальным трансфер результатов в задачу оптимальной остановки. Статья также использует ограничивающее предположение компактного пространства состояний и Липшиц-непрерывность переходных плотностей с константой Липшица, независимой от размерности d.

1.1.2 Метрики сложности

Оказывается, что критерий (2) придаёт слишком большое значение размерности d с одной стороны, и с другой стороны, слишком мало – зависимости от ε. С подобным

критерием алгоритм, обладающей сложностью $d^2\exp\left(\varepsilon^{-1}/\ln\ln ...\ln\varepsilon^{-1}\right)$, решаемый, в то время, как другой со сложностью $2^d/\varepsilon$ – нет; при этом реальный запуск алгоритма 1 на практике кажется практически невозможным даже с $d = 1$. Мы предлагаем другой критерий оценки сложности.

Определение 3. *Для алгоритма с вычислительной сложностью $\mathcal{C}(\varepsilon, d)$ число*

$$\Gamma_{\mathcal{C}} := \limsup_{d\to\infty} \limsup_{\varepsilon\to 0} \frac{\ln \mathcal{C}(\varepsilon, d)}{d\ln(1/\varepsilon)}.$$

называется индексом ST-решаемости (semitractability index).

Определение 4. *Задача называется ST-решаемой, если существует алгоритм, решающий её и обладающий индексом $\Gamma_{\mathcal{C}} = 0$.*

Заметим, что этот критерий лучше учитывает сложность вроде $1/\varepsilon^{poly(d)}$, делая возможным сравнение различных алгоритмов Монте-Карло для оптимальной остановки и оптимального управления.

1.1.3 Алгоритм WSM

Алгоритм взвешенного Монте-Карло (Weighted Stochastic Mesh, WSM) – это вдохновлённый [13] метод решения задачи оптимальной остановки в дискретном времени, и от оригинального он отличается специальным выбором весов и наличием обрезания маловероятных значений процесса. Для начала определим процесс огибающих Снелла:

$$U_l = U_l(Z_l) := \sup_{\tau\in\mathcal{T}_{l,L}} \mathbb{E}\left[g_\tau(Z_\tau) \mid \mathcal{F}_l\right], \quad l = 0, ..L,$$

где $\mathcal{T}_{l,L}$ – множество времён остановок, принимающих значения в $\{l, .., L\}$. Огибающая Снелла удовлетворяет принципу динамического программирования, следовательно, мы можем вычислить U_l используя обратную индукцию:

$$\begin{aligned} &U_L(Z_L) = g_L(Z_L),\\ &U_l(Z_l) = \max\left\{g_l(Z_l),\ \mathbb{E}\left[U_{l+1}(Z_{l+1}) \mid Z_l\right]\right\}, \quad l = 0, .., L-1. \end{aligned}$$

Для технических целей анализа мы задаём уровень обрезания $R > 0$ и определяем обрезанную версию обратной индукции:

$$\widetilde{U}_L(Z_L) = g_L(Z_L), \tag{3}$$

$$\widetilde{U}_l(Z_l) = \max\left\{g_l(Z_l),\ \mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}) \mid Z_l\right]\right\}\cdot \mathbb{1}_{B_R}(Z_l), \quad l = 0, .., L-1, \tag{4}$$

где $\mathbb{1}_{B_R}$ – индикаторная функция евклидова шара радиуса R и центром в точке O в $\mathbb{R}^d$. Так, значения обращаются в ноль, когда процесс выходит за пределы B_R. Мы семплируем N независимых траекторий $(Z_l^{(n)})_{l\in\mathbb{L}}$ с начальным значением $Z_0^{(n)} = x_0, n = 1, .., N$ с помощью переходной плотности $p(y|x)$. Для оценки условных матожиданий мы используем следующую аппроксимацию:

$$\mathbb{E}\left[\widetilde{U}_{l+1}(Z_{l+1}) \mid Z_l = x\right] \approx \sum_{n=1}^{N} \widetilde{U}_{l+1}\left(Z_{l+1}^{(n)}\right) \frac{p\left(Z_{l+1}^{(n)} \mid x\right)}{\sum_{m=1}^{N} p\left(Z_{l+1}^{(n)} \mid Z_l^{(m)}\right)}. \tag{5}$$

Подытожив, приходим к алгоритму WSM:

1. Симулировать N независимых траекторий $(Z_l^{(1)})_{l\in\mathcal{L}}, .., (Z_l^{(N)})_{l\in\mathcal{L}}$;

2. Задать $\overline{U}_L(Z_L^{(n)}) = g_L(Z_L^{(n)})$ для $n = 1, .., N$;

3. Для $l = L-1, .., 1$ вычислить $\overline{U}_l(Z_l^{(n)})$ для всех $n = 1, .., N$, используя (4) и (5) для аппроксимации условных матожиданий;

4. Вычислить
$$\overline{U}_0(x_0) = \max\left\{g_0(x_0)\ ,\ \frac{1}{N}\sum_{n=1}^{N}\overline{U}_1^{(n)}\left(Z_1^{(n)}\right)\right\}.$$

Ещё один факт, который необходимо заметить, это то, что один шаг обратной индукции (4) с аппроксимацией (5) занимает N^2c_*, где c_* – цена одного умножения. Таким образом, общая вычислительная сложность алгоритма есть c_*N^2L и при условии $c_* \ll c_f^{(d)}$ на цену одного вычисления переходной плотности, она ограничена сверху $c_f^{(d)}N^2L$.

1.1.4 Основные результаты

Используя оценки из литературы, мы вычислили индекс $\Gamma_{\mathcal{C}}$ для популярных методов Лонгштаффа-Шварца [44] и Квантизационного Дерева(Quantization Tree Method, QTM) [4], результаты для дискретного и непрерывного времени можно видеть в таблице ниже. Для WSM мы получили следующие два больших результата.

Теорема 1. *(Proposition 2.5 in [6]) Предположим, что выполнены следующие условия:*

1.
$$\max_{0\le l\le L} g_l(x) \le c_g(1+\|x\|_2), \quad x\in\mathbb{R}^d;$$

2.
$$\mathbb{E}\left[\max_{l\le l'\le L}|Z_{l'}| \ \middle|\ Z_l = x\right] \le c_Z(1+\|x\|_2), \quad x\in\mathbb{R}^d;$$

3. *Существуют $\kappa, \alpha > 0$, такие, чтодля всех $l = 1, .., L$ переходная плотность за l шагов удовлетворяет*
$$0 < p_l(y|x) \le \frac{\kappa}{(2\pi\alpha L)^{d/2}} e^{-\frac{\|x-y\|_2^2}{2\alpha l}}.$$

Тогда сложность WSM-алгоритма ограничивается сверху выражением

$$\mathcal{C}(\varepsilon, d) = c_1\alpha^2 c_g^4\kappa^2 c_f^{(d)} c_2^d L^{d+7}\varepsilon^{-4} \times \ln^{d+2}\left[\frac{L\left(1+c_Z+c_Z\|x_0\|_2\right)e^{\frac{c_Z\sqrt{\alpha L}}{1+c_Z+c_Z\|x_0\|_2}}2^{3/4}\left(c_g\kappa\vee 1\right)}{\varepsilon}\right].$$

Следствие 2. *(Corollary 2.6 in [6]) Задача оптимальной остановки в дискретном времени в предположении Теоремы 1 является ST-решаемой, если сложность вычисления переходной плотности $c_f^{(d)}$ не более, чем полиномиально зависит от d.*

Мы получили также побочный результат, говорящий о том, что даже если мы не знаем переходной плотности, но можем её достаточно хорошо аппроксимировать, то результат остаётся верным с немного другими константами. В частности, мы получаем конечный ST-индекс, если аппроксимирующая последовательность p^n удовлетворяет

$$\left|\frac{p^n(y|z) - p(y|z)}{p^n(y|z)}\right| \lesssim \frac{\left(1 + \|y - x_0\|_2^m + \|z - x_0\|_2^m\right)^n}{n!}, \quad y, z \in B_{R_n}$$

для некоторых $m \in \mathbb{Z}_{>0}$ и подходящей последовательности $R_n \to \infty$ при $n \to \infty$.

Для решения задачи в непрерывном времени мы сначала вводим схему дискретизации, основанную на методе Эйлера сравномерным шагом по времени h (детальнее см. [6]). Это приводит нас к уже рассмотренной задаче в дискретном времени. На самом же деле, Теорема работает и в более общих случаях, схема Эйлера – это лишь один конкретный подход, которого уже оказывается достаточно.

Теорема 3. *(Proposition 3.4 in [6]) Пусть верны следующие условия:*

1. $$\max_{0 \le t \le T} g_t(x) \le c_g(1 + \|x\|_2), \quad x \in \mathbb{R}^d;$$

2. $$\mathbb{E}\left[\max_{l \le l' \le L} \left|\overline{X}_{l'h}\right| \;\middle|\; \overline{X}_{lh} = x\right] \le c_{\overline{X}}(1 + \|x\|_2), \quad x \in \mathbb{R}^d;$$

3. *Существуют константы $\overline{\kappa}, \overline{\alpha} > 0$ такие, что для всех $l = 1, .., L$ переходная плотность за l шагов процесса $(\overline{X}_{lh})_{l \in \mathcal{L}}$ удовлетворяет*

$$0 < \overline{p}_{lh}(y|x) \le \frac{\overline{\kappa}}{(2\pi\overline{\alpha}lh)^{d/2}} e^{-\frac{\|x-y\|_2^2}{2\overline{\alpha}lh}}.$$

Тогда вычислительная сложность решения полученной задачи в дискретном времени ограничена сверху выражением

$$\mathcal{C}(\varepsilon, d) = c_1\overline{\alpha}^2 c_g^4 \overline{\kappa}^2 c_f^{(d)} c_2^d \frac{T^{d+7}}{h^{d+5}} \varepsilon^{-4} \times \ln^{d+2}\left[\frac{(T/h)\left(1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2\right) e^{\frac{c_{\overline{X}}\sqrt{\overline{\alpha}T}}{1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2}} 2^{3/4}\left(c_g\overline{\kappa} \vee 1\right)}{\varepsilon}\right]$$

и, в свою очередь, вычислительная сложность задачи в непрерывном времени ограничена сверху

$$\mathcal{C}^\star(\varepsilon, d) = c_1\overline{\alpha}^2 c_g^4 \overline{\kappa}^2 c_f^{(d)} c_2^d \frac{T^{d+7}}{\varepsilon^{2d+14}} \times \ln^{d+2}\left[\frac{T\left(1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2\right) e^{\frac{c_{\overline{X}}\sqrt{\alpha T}}{1 + c_{\overline{X}} + c_{\overline{X}}\|x_0\|_2}} 2^{3/4}\left(c_g\overline{\kappa} \vee 1\right)}{\varepsilon}\right].$$

Следствие 4. *Для задачи оптимальной остановки в непрерывном времени WSM-алгоритм с дискретизацией, удовлетворяющей предположениям Теоремы 3, имеет ST-индекс $\Gamma_{\mathcal{C}^\star} = 2$.*

Сравнительная таблица полученных ST-индексов приведена в нашей статье [6] и помещена ниже.

Setting \ Algorithm	LS	WSM	QTM
Discr. time	$3/\alpha$	0	2
Cont. time	∞	2	6

Таблица 1: Таблица ST-индексов для методов Лонгштаффа-Шварца(LS), WSM и метода Квантизационного Деревва (QTM) вычисленных в статье.

1.1.5 Численные эксперименты

В экспериментах мы демонстрируем работу WSM-алгоритма в задаче оптимальной остановки в непрерывном времени. нижняя оценка для WSM-метода получена с использованием субоптимального правила остановки, вычисленного на независимом множестве траекторий (тестовое множество. Это правило остановки может быть составлено с использованием любого алгоритма интерполяции, основанного на наблюдениях из тренировочного множества. Наибыстрейший и наипростейший способ, дающий хорошие результаты – метод ближайших соседей, в наших экспериментах мы задавали число ближайших соседей равное 500.

Американский пут-опцион на один актив

Чтобы проиллюстрировать работу WSM-алгоритма в непрерывном времени, мы рассматриваем задачу оценки американского пут-опциона на один актив, эволюция цен которого задаётся моделью геометрического Броуновского движения

$$X_t = X_0 e^{\sigma W_t + (r - \sigma/2)t}$$

с r, безрисковой процентной ставкой (предполагается константной), и σ – константной волатильностью. Функция выплат для подобного опциона задаётся как

$$g(x) = \max(K - x, 0)$$

и честная цена опциона определяется

$$U_0 = \sup_{\tau \in \mathcal{T}_{[0,T]}} \mathbb{E}\left[e^{-r\tau} g(X_\tau)\right],$$

для которой нет решения в виде формулы, но существуют численные методы, дающие достаточно точные приближения. Мы использовали параметры $r = 0.08, \sigma = 0.20, K = X_0 = 100, T = 3$. Точное приближение U_0 в данном конкретном случае получено в статье [36] и составляет 6.9320. На Рис. 1 мы показываем нижние оценки, полученные WSM, LS и VF (регрессионный метод Тситсиклиса и Ван Роя [67]) в зависимости от количества возможностей остановки L, равномерной сеткой на $[0, T]$ (больше L означает более плотную сетку дискретизации). Как можно заметить, нижняя оценка, полученная WSM, гораздо более устойчива при росте L, в то время как LS и VF сильно ухудшаются и требуют дополнительных базисных функций, чтобы компенсировать этот эффект.

Американский макс-колл опцион на 5 активов

Рассматривается модель с $d = 5$ активами, где каждый актив обладает дивидендом δ. Динамика задаётся уравнением

$$dX_t^k = (r - \delta)X_t^k dt + \sigma X_t^k dW_t^k, \quad k = 1, .., d,$$

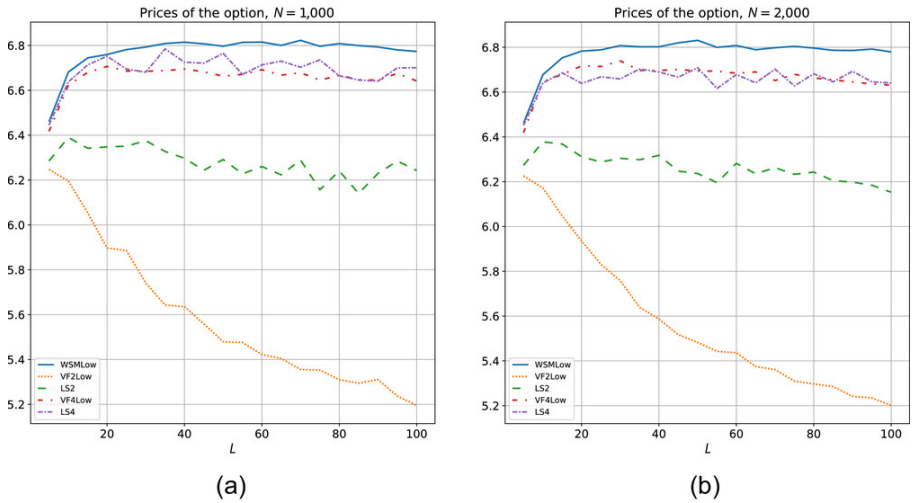

Рис. 1: Нижние оценки цены одномерного американского пут-опциона, полученные разными методами и равномерной сеткой дискретизации по времени с $t_k = kT/L, k = 0, .., L$ – возможными моментами для остановки. Число тренировочных траекторий равно $N_{train} = 1000$(a) и $N_{train} = 2000$(b) и число тестовых траекторий, использованное для построения нижней оценки составляет $N_{test} = 20000$, одинаково в обоих случаях. В LS и VF использовались базисы из полиномов степеней до 2 и до 4 (согласно легенде). График из [6].

где W_t^k – независимые одномерные Броуновские движения. Параметры заданы $r = 0.05, \delta = 0.1, \sigma = 0.2$. Как и ранее, владелец может исполнить опцион в момент $t \in [0, T]$ с $T = 3$ и получить выплату

$$g(X_t) = max\left(\max\left(X_t^1, .., X_t^d\right) - K,\ 0\right).$$

Мы используем WSM и LS алгоритмы (с базисом из полиномов степени не выше 2), чтобы получить нижнюю оценку. Результаты для различных L представлены на Рис. 2. Цена опциона должна увеличиваться при росте числа возможных точек остановки, следовательно, LS-алгоритм ощутимо теряет в качестве. WSM, с другой стороны, имеет возрастающую нижнюю оценку, демонстрирующую, что WSM-алгоритм работает гораздо точнее, чем LS.

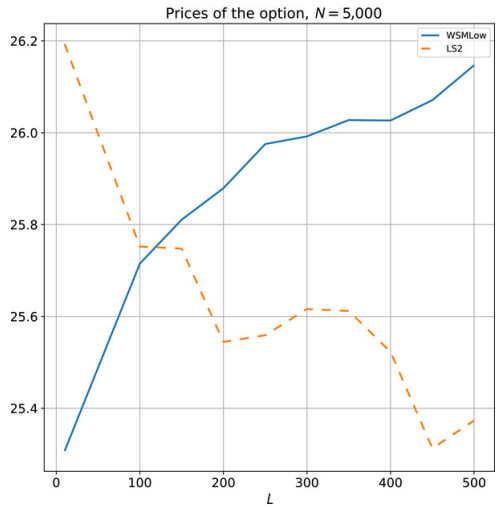

Рис. 2: Нижние оценки цены 5-мерного американского пут-опциона, полученные с использованием равномерной сетки с возможностями остановки $t_k = kT/L, k = 0, .., L$. Число траекторий для обучения равно $N_{train} = 2000$, а тестовых $N_{test} = 5000$. График из [6].

1.2 Неасимптотический анализ линейной двумасштабной схемы стохастической аппроксимации с марковским шумом

Результаты этой главы опубликованы в статье [35].

1.2.1 Введение

Схема TD-learning, основанная на классической (линейной) стохастической аппроксимации, становится неадекватной в условиях парадигмы off-policy в RL, где данные получаются с помощью *поведенческой политики*, которая отличается от оцениваемой [3, 66]. Чтобы обойти эту проблему, в [59, 60] был предложен метод градиентного TD (GTD) и TD с градиентной коррекцией (TDC). Эти методы в сущности являются схемами линейной стохастической аппроксимации с двумя масштабами, впервые представленными в работе [11]:

$$\theta_{k+1} = \theta_k + \beta_k\{\widetilde{b}_1(X_{k+1}) - \widetilde{A}_{11}(X_{k+1})\theta_k - \widetilde{A}_{12}(X_{k+1})w_k\}, \tag{6}$$

$$w_{k+1} = w_k + \gamma_k\{\widetilde{b}_2(X_{k+1}) - \widetilde{A}_{21}(X_{k+1})\theta_k - \widetilde{A}_{22}(X_{k+1})w_k\}. \tag{7}$$

Рекурсия включает в себя два набора переменных, $\theta_k \in \mathbb{R}^{d_\theta}$, $w_k \in \mathbb{R}^{d_w}$, чьи обновления связаны. В вышеприведённом $\widetilde{b}_i(x)$, $\widetilde{A}_{ij}(x)$ – измеримые функции на измеримом пространстве X, принимающие в качестве значения вектор или матрицу и случайная последовательность $(X_k)_{k\geq 0}, X_k \in \mathrm{X}$ предполагается марковской цепью. Мы также задаём скаляры $\gamma_k, \beta_k > 0$ – величины шага обновления. Вышеприведённая схема стохастической аппроксимации, как принято говорить, имеет два масштаба, так как величины шагов удовлетворяют $\lim_{k\to\infty} \beta_k/\gamma_k < 1$; таким образом, w_k обновляется быстрее. На самом деле, w_k – is a 'следящий' член, аппроксимирующий решение линейной системы, возникающей при фиксированном $\theta = \theta_k$.

Наша цель – предложить неасимптотическую оценку ошибки с улучшенными скоростями сходимости для двумасштабной схемы (6),(7). Сходимость почти наверное приведённой схемы было установлено в работах [11, 62, 63, 12] среди прочих, а также в [38, 48] приведены асимптотические скорости сходимости в линейном случае. Тем не менее, неасимптотические (по времени) оценки ошибки для двумасштабной схемы до недавнего момента не получены. В случае мартингального шума статья [42] представила впервые неасимптотический анализ метода GTD, который был улучшен далее в [21, 19]. В отличие от нашего анализа, они анализируют модифицированную схему, в которую включён шаг проекции, а приведённые оценки – оценки с высокой вероятностью. Что касается Марковского шума, в [33] изучали неасимптотические оценки ошибки при константных шагах; [74] и [22] предоставили похожий анализ для более общей динамики изменения шага. Очень важно отметить, что при однородном мартингальном шуме, асимптотические скорости сходимости (6), (7) без проекций, как показано в [38, Theorem 2.6], составляют порядка $\mathbb{E}\left[|\theta_k - \theta^\star|^2\right] = \mathcal{O}(\beta_k), \mathbb{E}\left[\left|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\right|^2\right] = \mathcal{O}(\gamma_k)$, где $\theta^\star$ – стационарная точка схемы стохастической аппроксимации. Всё же, это скорость не достигается ни в одной из вышеприведённых работ, кроме [19]. Было открытой задачей понять, совпадают ли асимптотическая и неасимптотическая скорости сходимости в случае Марковского шума и линейной схемы без проекции.

1.2.2 Основные результаты

Исследуется линейная двумасштабная линейная схема стохастической аппроксимации в форме, эквивалентной (6) и (7):

$$\theta_{k+1} = \theta_k + \beta_k(b_1 - A_{11}\theta_k - A_{12}w_k + V_{k+1}), \tag{8}$$
$$w_{k+1} = w_k + \gamma_k(b_2 - A_{21}\theta_k - A_{22}w_k + W_{k+1}), \tag{9}$$

где $b_i := \lim_{k\to\infty} \mathbb{E}\left[\widetilde{b}_i(X_k)\right]$, $A_{ij} := \lim_{k\to\infty} \mathbb{E}\left[\widetilde{A}_{ij}(X_k)\right]$ (пределы существуют почти наверное, так как $(X_k)_{k\geq 0}$ – эргодическая цепь Маркова). Шумы V_{k+1}, W_{k+1} задаются как

$$\begin{aligned} V_{k+1} &:= \widetilde{b}_1(X_{k+1}) - b_1 - (\widetilde{A}_{11}(X_{k+1}) - A_{11})\theta_k - (\widetilde{A}_{12}(X_{k+1}) - A_{12})w_k, \\ W_{k+1} &:= \widetilde{b}_2(X_{k+1}) - b_2 - (\widetilde{A}_{21}(X_{k+1}) - A_{21})\theta_k - (\widetilde{A}_{22}(X_{k+1}) - A_{22})w_k. \end{aligned} \tag{10}$$

Цель рекурсии (8), (9) состоит в том, чтобы найти стационарное решение $(\theta^\star, w^\star)$ системы линейных уравнений

$$A_{11}\theta + A_{12}w = b_1, \quad A_{21}\theta + A_{22}w = b_2. \tag{11}$$

Мы заинтересованы в случае, где $(\theta^\star, w^\star)$ – единственное решение и, таким образом, представимо в виде

$$\theta^\star = \Delta^{-1}(b_1 - A_{12}A_{22}^{-1}b_2), \quad w^\star = A_{22}^{-1}(b_2 - A_{21}\theta^\star), \tag{12}$$

где $\Delta := A_{11} - A_{12}A_{22}^{-1}A_{21}$.

Для анализа сходимости $(\theta_k, w_k)_{k\geq 0}$ схем (8), (9) к $(\theta^\star, w^\star)$ нужно несколько типичных для задачи технических предположений [38].

A 1. Матрицы $-A_{22}$ и $-\Delta = -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)$ являются *гурвицевыми*.

A 2. $(\gamma_k)_{k\geq 0}$, $(\beta_k)_{k\geq 0}$ – невозрастающие последовательности положительных чисел, удовлетворяющих следующим условиям:

1. Существует константа κ, такая, что для $k \in \mathbb{N}$, верно $\beta_k/\gamma_k \leq \kappa$.
2. Для всех $k \in \mathbb{N}$ верно

$$\gamma_k/\gamma_{k+1} \leq 1+(a_{22}/8)\gamma_{k+1}, \;\; \beta_k/\beta_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}, \;\; \gamma_k/\gamma_{k+1} \leq 1+(a_\Delta/16)\beta_{k+1}. \tag{13}$$

Наши условия на размер шага похожи на [38, Assumption 2.3, 2.5]. Они позволяют рассматривать убывающие, кусочно константные и константные динамики шага, обычно исследуемые в литературе. Например, популярный выбор шага, удовлетворяющий A2, – это

$$\beta_k = c^\beta/(k + k_0^\beta), \quad \gamma_k = c^\gamma/(k + k_0^\gamma)^{2/3} \tag{14}$$

с некоторыми константами c^β, c^γ, k_0^γ, k_0^β, как, к примеру, предложено в [21, Remark 9]; либо константные шаги $\beta_k = \beta, \gamma_k = \gamma$ или кусочно константные, как в [33].

Мы показываем новые результаты о скорости сходимости (8), (9), в зависимости от типа шумов V_{k+1}, W_{k+1}. Для обсуждения этих случаев, определим σ-поле, порождённое двумасштабной схемой и изначальную ошибку соответственно как

$$\mathcal{F}_k := \sigma\{\theta_0, w_0, X_1, X_2, ..., X_k\}, \quad \mathrm{V}_0 := \mathbb{E}\left[\|\theta^0 - \theta^\star\|^2 + \|w^0 - w^\star\|^2\right]. \tag{15}$$

Наши главные результаты представлены для двух типов шума.

Мартингальный шум. Рассмотрим более простой случай, когда X_k – это i.i.d.-последовательность из распределения, такого, что b_i, A_{ij} – матожидания случайных величин $\widetilde{b}_i(X_k), \widetilde{A}_{ij}(X_k)$, которые имеют конечный второй момент. Это даёт, что последовательности $(V_{k+1})_{k\in\mathbb{N}}, (W_{k+1})_{k\in\mathbb{N}}$ являются последовательностями мартингальных разностей. Последнее может быть верно и в более общих случаях.

A 3. Шумовые члены имеет нулевое матожидание при условии $\mathcal{F}_k$: $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] = \mathbb{E}^{\mathcal{F}_k}[W_{k+1}] = 0$.

A 4. Существуют константы m_W, m_V такие, что

$$\begin{aligned}\|\mathbb{E}\left[V_{k+1}V_{k+1}^\top\right]\| &\le m_V(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|),\\ \|\mathbb{E}\left[W_{k+1}W_{k+1}^\top\right]\| &\le m_W(1 + \|\mathbb{E}\left[\theta_k\theta_k^\top\right]\| + \|\mathbb{E}\left[w_k w_k^\top\right]\|)\,.\end{aligned} \tag{16}$$

Теорема 5. *Пусть верно A1–4 и для всех $k \in \mathbb{N}$ имеем $\gamma_k \in [0, \gamma_\infty^{\mathsf{mtg}}]$, $\beta_k \in [0, \beta_\infty^{\mathsf{mtg}}]$ и $\kappa \in [0, \kappa_\infty]$, где $\gamma_\infty^{\mathsf{mtg}}, \beta_\infty^{\mathsf{mtg}}, \kappa_\infty$ – конкретные константы. Тогда*

$$\mathbb{E}\left[\|\theta_k - \theta^*\|^2\right] \le d_\theta\left\{\mathrm{C}_0^{\tilde{\theta},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\tilde{\theta},\mathsf{mtg}}\beta_k\right\} \tag{17}$$

$$\mathbb{E}\left[\left\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\right\|^2\right] \le d_w\Big\{\mathrm{C}_0^{\widehat{w},\mathsf{mtg}}\prod_{\ell=0}^{k-1}\Big(1 - \beta_\ell\frac{a_\Delta}{4}\Big)\mathrm{V}_0 + \mathrm{C}_1^{\widehat{w},\mathsf{mtg}}\gamma_k\Big\} \tag{18}$$

Точные константы представлены в статье.

Марковский шум. Рассмотрим последовательность $(X_k)_{k\ge0}$ семплов из внешней Марковской цепи X с переходным ядром $\mathrm{P} : \mathsf{X}\times\mathsf{X} \to \mathbb{R}_+$. Для любой измеримой функции f имеем

$$\mathbb{E}^{\mathcal{F}_k}[f(X_{k+1})] = \mathrm{P}\,f(X_k) = \int_{\mathsf{X}} f(x)\,\mathrm{P}(X_k, \mathrm{d}x).$$

B 1. Марковское ядро P имеет единственное инвариантное распределение $\mu : \mathsf{X} \to \mathbb{R}_+$. Более того, цепь является несократимой и апериодичной.

Заметим, что

$$b_i = \int_{\mathsf{X}} \widetilde{b}_i(x)\,\mu(\mathrm{d}x), \quad A_{ij} = \int_{\mathsf{X}} \widetilde{A}_{ij}(x)\,\mu(\mathrm{d}x),\ i,j = 1,2.$$

Мы показываем, что схема (6), (7) сходится к единственной стационарной точке, определённой в (12). Важным условием, позволяющим наш анализ, является существование решения уравнений Пуассона.

B 2. Для $i,j = 1,2$ и $\widetilde{b}_i(x), \widetilde{A}_{ij}(x)$, существуют измеримые функции $\widehat{b}_i(x), \widehat{A}_{ij}(x)$, удовлетворяющие

$$\widetilde{b}_i(x) - b_i = \widehat{b}_i(x) - \mathrm{P}\,\widehat{b}_i(x), \quad \widetilde{A}_{ij}(x) - A_{ij} = \widehat{A}_{ij}(x) - \mathrm{P}\,\widehat{A}_{ij}(x) \tag{19}$$

для всех $x \in \mathsf{X}$ и b_i, A_{ij} – средние $\widetilde{b}_i(x), \widetilde{A}_{ij}(x)$, посчитанные по стационарному распределению μ.

Вышеприведённое предположение гарантировано B1 и некоторыми условиями регулярности, см. [23, Section 21.2]. Более того,

B 3. При B2, функции $\widehat{b}_i(x), \widehat{A}_{ij}(x)$ равномерно ограничены: для $i, j = 1, 2$, $x \in \mathsf{X}$,

$$\|\widehat{b}_i(x)\| \le \overline{\mathrm{b}}, \ \|\widehat{A}_{ij}(x)\| \le \overline{\mathrm{A}}. \tag{20}$$

B 4. Существует константа ρ_0 такая, что для всех $k \ge 1$ имеем $\gamma_{k-1}^2 \le \rho_0 \beta_k$.

Чтобы выполнилось B3, заметим, что оценки $\overline{\mathrm{b}}, \overline{\mathrm{A}}$ зависят от времени смешивания цепи $(X_k)_{k\ge 0}$ и равномерной оценки на $\widetilde{b}_i(\cdot), \widetilde{A}_{ij}(\cdot)$. В контексте RL последнее выполнено в случае ограниченных наград и векторов признаков. В действительности B3 влечёт A4. Тем не менее, B4 задаёт более сильные ограничения на размер шага. Последнее может быть также удовлетворено с помощью (14). Основная сложность при анализе случая с Марковским шумом состоит в смещённости шумового члена, так как $\mathbb{E}^{\mathcal{F}_k}[V_{k+1}] \neq 0$, $\mathbb{E}^{\mathcal{F}_k}[W_{k+1}] \neq 0$.

Теорема 6. *Положим A1–2, B1–4 верными и пусть для всех $k \in \mathbb{N}$ размеры шага $\beta_k \in (0, \beta_\infty^{\mathsf{mark}}]$, $\gamma_k \in (0, \gamma_\infty^{\mathsf{mark}}]$, $\kappa \le \kappa_\infty$, где $\beta_\infty^{\mathsf{mark}}$, $\gamma_\infty^{\mathsf{mark}}$, κ_∞ – определённые константы. Тогда*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] \le d_\theta \left\{ \mathrm{C}_0^{\tilde{\theta},\mathsf{mark}} \prod_{\ell=0}^{k-1} \Big(1 - \beta_\ell \frac{a_\Delta}{8}\Big)(1 + \mathrm{V}_0) + \mathrm{C}_1^{\tilde{\theta},\mathsf{mark}} \beta_k \right\}, \tag{21}$$

$$\mathbb{E}\left[\|w_k - A_{22}^{-1}(b_2 - A_{21}\theta_k)\|^2\right] \le d_w \left\{ \mathrm{C}_0^{\widehat{w},\mathsf{mark}} \prod_{\ell=0}^{k-1} \Big(1 - \beta_\ell \frac{a_\Delta}{8}\Big)(1 + \mathrm{V}_0) + \mathrm{C}_1^{\widehat{w},\mathsf{mark}} \gamma_k \right\}. \tag{22}$$

Точные выражения для констант приведены в статье.

В то время, как Теорема 6 ослабляет предположение мартингальных разностей A4 в Теореме 5, можно заметить, что результаты не обобщают в полной мере Теорему 5 в силу дополнительных B3, B4. В частности, в предположении мартингального шума, сходимость схемы требует от шума только ограниченного *второго момента*, но в Марковском случае требуется равномерная ограниченность.

Верхние оценки Теоремы 5 и 6 состоят из двух частей: первый член – 'уходящая' ошибка с произведением $\prod_{i=0}^{k-1}(1 - \beta_i a_\Delta/8)$, убывающим к нулю со скоростью $o(1/k^c)$ для некоторого $c > 1$ при подходящем выборе размера шагов, как, например, в (14); второй член – 'стационарная' ошибка. Заметим, что 'стационарная' ошибка в итерациях θ_k, w_k ведёт себя по-разному. Если в качестве примера взять (14), то стационарная ошибка медленно обновляющейся итерации θ_k составляет $\mathcal{O}(1/k)$, тогда как ошибка быстро обновляющейся итерации w_k равна $\mathcal{O}(1/k^{\frac{2}{3}})$. Более того, похожие оценки верны *сразу* для мартингального и Марковского шума.

Сравнение с похожими работами. Наши результаты улучшают уже полученные в анализе линейных схем стохастической аппроксимации. В предположении мартингального шума (Теорема 5), самая близкая к нам статья [19] улучшала результаты [21] и получила те же скорости сходимости (с высокой вероятностью), что приведены в Теореме 5 что ещё раз подтверждает точность полученных нами оценок. Их оценки также имеют сублинейную зависимость от размерности d_θ, d_w. Тем не менее, их алгоритм включает в себя специальный шаг разреженной проекции и полученные оценки

верны только для достаточно больших k. Эти ограничения были ликвидированы в нашем анализе.

В Марковском случае (Теорема 6) ближе всего к нашей стоят работы [22, 33, 74]. В частности, в [33] анализировалась схема с константными шагами и было показано, что стационарная ошибка в обеих итерациях θ_k, w_k составляет $\mathcal{O}(\gamma^2/\beta)$. В [74] анализировали алгоритм TDC с проекционным шагом и показали, что стационарная ошибка в итерациях θ_k равна $\mathcal{O}(1/k^{\frac{2}{3}})$, если использованы величины шагов из (14). Статья [22] анализировала схему с убывающими шагами, и для итераций θ_k, w_k была получена стационарная ошибка $\mathcal{O}(1/k^{\frac{2}{3}})$. Что особенно интересно, вышеприведённые работы не получают быстрой сходимости Теоремы 6, т.е., порядка $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = \mathcal{O}(1/k)$. Одна из причин субоптимальности состоит в том, что анализ основывался на построении одной функции Ляпунова для ошибок в θ_k и w_k. Наш анализ, напротив, основывается на связанных неравенствах, позволяющих получить точные оценки для θ_k, w_k.

Наш последний результат – нижняя оценка, построенная для демонстрации точности нашего анализа в Теореме 5, 6. Основная идея – выписать явное выражение для $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$. Мы требуем следующее техническое предположение.

A 5. Существуют матрицы $\Sigma^{11}, \Sigma^{12}, \Sigma^{22}$, и константа $m_{VW}^{\mathsf{exp}} \geq 0$ такие, что для всех $j \in \mathbb{N}$ выполняется

$$\|\mathbb{E}\left[V_j V_j^\top\right] - \Sigma^{11}\| \vee \|\mathbb{E}\left[W_j W_j^\top\right] - \Sigma^{22}\| \vee \|\mathbb{E}\left[V_j W_j^\top\right] - \Sigma^{12}\| \leq m_{VW}^{\mathsf{exp}}(\|\mathbb{E}\left[\tilde{\theta}_k \tilde{\theta}_k^\top\right]\| + \|\mathbb{E}\left[\tilde{w}_k \tilde{w}_k^\top\right]\|).$$

Заметим, что A5 влечёт A4 и, следовательно, представляет собой более сильное предположение. В результате получается

Теорема 7. *Пусть верны A1–3, A5 и для всех $k \in \mathbb{N}$ имеем $\gamma_k \in [0, \gamma_\infty^{\mathsf{mtg}}]$, $\beta_k \in [0, \beta_\infty^{\mathsf{exp}}]$ и $\kappa \in [0, \kappa_\infty^{\mathsf{exp}}]$, где $\gamma_\infty^{\mathsf{mtg}}, \beta_\infty^{\mathsf{exp}}, \kappa_\infty^{\mathsf{exp}}$ – конкретные константы, приведённые в статье. Тогда для каждого $k \geq k_0^{\mathsf{exp}} := \min\{\ell : \sum_{j=0}^{\ell-1} \beta_j \geq \log(2)/(2\|\Delta\|)\}$ верно разложение*

$$\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right] = I_k + J_k. \tag{23}$$

Ведущий член I_k задаётся явным выражением

$$I_k := \textstyle\sum_{j=0}^k \beta_j^2 \operatorname{Tr}\left(\prod_{\ell=j+1}^k (\mathrm{I} - \beta_\ell \Delta)\, \Sigma \left\{\prod_{\ell=j+1}^k (\mathrm{I} - \beta_\ell \Delta)\right\}^\top\right),$$

где $\Sigma := \Sigma^{11} + A_{12}A_{22}^{-1}\Sigma^{22}A_{22}^{-\top}A_{12}^\top + \Sigma^{12}A_{22}^{-\top}A_{12}^\top + A_{12}A_{22}^{-1}\Sigma^{21}$. Также верно неравенство

$$\mathrm{C}_3^{\mathsf{exp}} \operatorname{Tr}(\Sigma) \leq \frac{I_k}{\beta_k} \leq \mathrm{C}_4^{\mathsf{exp}} \operatorname{Tr}(\Sigma), \tag{24}$$

и J_k ограничены как

$$|J_k| \leq \mathrm{C}_0^{\mathsf{exp}} \prod_{\ell=0}^{k-1} \left(1 - \frac{a_\Delta}{4}\beta_\ell\right) \mathrm{V}_0 + \mathrm{C}_1^{\mathsf{exp}} \beta_k \left(\gamma_k + \frac{\beta_k}{\gamma_k}\right), \tag{25}$$

где V_0 дано (15). Все константы $\mathrm{C}_0^{\mathsf{exp}}, \mathrm{C}_1^{\mathsf{exp}}, \mathrm{C}_3^{\mathsf{exp}}, \mathrm{C}_4^{\mathsf{exp}}$ приведены в статье и не зависят от β_k, γ_k.

Заметим, что из (25) получаем, что ведущий член в J_k – это $\mathcal{O}(\beta_k\gamma_k + \frac{\beta_k^2}{\gamma_k})$. Таким образом, используя (24), получаем

$$|J_k|/I_k = \mathcal{O}\left(\gamma_k + \beta_k/\gamma_k\right).$$

Если $\lim_{k\to\infty} \beta_k/\gamma_k = 0$, имеем $\lim_{k\to\infty} |J_k|/I_k = 0$. Совмещение (23), (24) показывает, что ожидаемая ошибка $\mathbb{E}\left[\|\theta_k - \theta^\star\|^2\right]$ ограничена снизу $\Omega(\beta_k)$.

Заметим, что A1–3, A5, требуемые в Теореме, влекут A1–A4, которые требуются в Теореме 5. Следовательно, вместе с (17) в Теореме 5, вышеприведённые наблюдения дают основания судить о *совпадении* нижней оценки скорости сходимости линейной схемы стохастической аппроксимации в случае мартингального шума.

1.3 Снижение дисперсии для методов Policy-Gradient с помощью минимизации эспирической дисперсии

Результаты этой главы опубликованы в [34].

1.4 Введение

В обучении с подкреплением (Reinforcement Learning, RL) методы policy-gradient составляют семейство градиентных алгоритмов, в основе которых лежит идея напрямую моделировать решающее правило(называемое обычно политикой) и использовать различные формулы для оценки градиента ожидаемой награды по параметрам политики [71, 61]. Самый прямой способ оценить градиент – использовать схему Монте-Карло, которая в результате приводит к самому первому методу REINFORCE [71]. Пусть задан Марковский процесс принятия решений (Markov Decision Process, MDP) $(\mathcal{S}, \mathcal{A}, R, \mathrm{P}, \Pi, \mu_0, \gamma)$ с конечным горизонтом T и дан класс политик

$$\Pi = \{\pi_\theta : \mathcal{S} \to \mathcal{P}(\mathcal{A}) \mid \theta \in \Theta\},$$

параметризованный $\theta \in \Theta \subset \mathbb{R}^D$, где $\mathcal{P}(\mathcal{A})$ – множество вероятностных распределений на пространстве действий $\mathcal{A}$. Мы будем опускать нижний индекс в π_θ, где возможно, для более краткой нотации; во всех случаях $\pi \in \Pi$. Оптимизационная задача для MDP задаётся как

$$\max\ J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t R(S_t, A_t)\right] \quad \text{по} \ \ \theta \in \Theta,$$

где горизонт T предполагается фиксированным. Заметим, что любая последовательность состояний, действий и наград может быт представлена как элемент X произведения измеримых пространств

$$(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T.$$

Пусть $\widetilde{\nabla} J|_{\theta'} : (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T \to \mathbb{R}^D$ является несмещённой оценкой градиента $\nabla_\theta J$ в точке $\theta = \theta'$. В этих нотациях градиентный алгоритм для максимизации $J(\theta)$, использующий оценку $\widetilde{\nabla} J$, записывается следующим образом:

$$\theta_{n+1} = \theta_n + \eta_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla} J|_{\theta_n}(X_n^{(k)}), \quad n = 1, 2, \ldots \tag{26}$$

при заданной последовательности положительных шагов η_n. В обозначении градиентных оценок мы будем опускать нижний индекс θ_n, если из контекста ясно, в какой точке вычисляется градиент. REINFORCE [71] является одним из примеров оценки градиента:

$$\widetilde{\nabla}^{\mathrm{reinf}} J :\ X \mapsto \sum_{t=0}^{T-1} \gamma^t G_t(X) \nabla_\theta \log \pi(A_t|S_t)$$

с

$$G_t(X) := \sum_{t'=t}^{T-1} \gamma^{t'-t} R_t,$$

где $R_t = R(S_t, A_t)$ и

$$X = [(S_0, A_0, R_0), .., (S_{T-1}, A_{T-1}, R_{T-1})]^\top .$$

Неизбежно такая оценка многомерного вектора градиента сильно зашумлена и обладает большой дисперсией [70], что делает задачу достаточно непростой. Для построения модификаций, улучшающих ситуацию, необходимо применять методы снижения дисперсии, чтобы сконструировать оценки, обладающие меньшей дисперсией и вычислительно более дешёвые, чем увеличение размера выборки Монте-Карло.

Основные достижения в этой области включают в себя метод actor-critic, предложенный в [37], и advantage actor-critic (A2C) [61], а также его асинхронная версия, A3C [46]. В целом, эти методы могут рассматриваться как модификация REINFORCE с помощью контрольных переменных, которые моделируются с помощью так называемого бейзлайна $b_\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, который может зависеть от действий (SA-бейзлайн), а может быть зависимым только от состояний $b_\phi : \mathcal{S} \to \mathbb{R}$ (S-бейзлайн); такой бейзлайн обычно выбирается из параметрического класса моделей с параметром ϕ. Скорректированная бейзлайном оценка имеет вид

$$\widetilde{\nabla}^{b_\phi}_\theta J : X \mapsto \sum_{t=0}^{T-1} \gamma^t (G_t - b_\phi(S_t, A_t)) \nabla_\theta \log \pi(A_t|S_t),$$

а градиентная схема для оптимизации по θ становится двумасштабной с новой итерацией обновления параметров бейзлайна, которые обновляются так, чтобы бейзлайн хорошо приближал функцию ценности и предсказывал сумму наград:

$$\theta_{n+1} = \theta_n + \alpha_n \frac{1}{K} \sum_{k=1}^{K} \widetilde{\nabla}^{b_\phi} J(X_n^{(k)}), \tag{27}$$

$$\phi_{n+1} = \phi_n - \beta_n \nabla_\phi V_{K,n}^{A2C}(\phi)|_{\phi_n}, \tag{28}$$

где

$$V_{K,n}^{A2C}(\phi) := \frac{1}{K} \sum_{k=1}^{K} \sum_{t=0}^{T-1} (G_t(X_n^{(k)}) - b_\phi(S_t^{(k)}))^2 \tag{29}$$

является целевым функционалом A2C, отражающим наше желание аппроксимировать функцию ценности, используя её зашумлённые оценки $(G_t(X_n^{(k)}))$, с помощью метода наименьших квадратов. Мотивация позади этого состоит в том, что если бейзлайн совпадает с функцией ценности, то дисперсия должна быть очень небольшой; на практике такая стратегия успешно работает [46].

Недавно подходы, связанные с бейзлайнами, получили новое развитие в свете развития методов глубинного обучения [47], очень исчерпывающий актуальный обзор можно найти в [27]. В течение двадцати лет было разработано много новых методов снижения дисперсии, включая подходы, основанные на использовании подвыборок, как Stochastic Variance-Reduced Policy-Gradient(SVRPG) [50, 73], и контрольных переменных [53], [32], [43], [68], [72]. Есть также некоторые необычные подходы, например, потраекторные(trajectory-wise) контрольные переменные [15], использующие контрольные переменные, основанные на будущих наградах и снижение дисперсии в средах с экзогенными переменными (input-driven environments) [45], где есть часть состояний, на которые агент никак не может повлиять. Помимо этого, в эргодическом

случае (когда MDP всегда стартует из состояния, данного стационарным распределением цепи) были получены как теоретические результаты [31], так и практические [17]. Важность критерия для снижения дисперсии хорошо известна в Монте-Карло и MCMC [54] и недавно этот вопрос возникал в [25], где был предложен метод Actor with Variance Estimated Critic (AVEC).

Если говорить о теоретической стороне, остаётся неясным, какое отношение имеет процедура в A2C к дисперсии градиента как таковой. Более того, эмпирические исследования дисперсии градиента достаточно редкие и доступны, в основном, только для искусственно сконструированных задач. В сообществе идут дискуссии о том, помогает ли вообще снижение дисперсии ускорению обучения [68]. В нашем исследовании мы пытаемся ответить на некоторые из этих вопросов и предлагаем более прямой подход, вдохновлённый минимизацией эмпирической дисперсией(Empirical Variance,EV), недавно изученной в [5]. Мы показываем, что предложенные EV-алгоритмы не только теоретически лучше обоснованы, но также способны работать лучше, чем классический алгоритм A2C. Отметим, что использование для обучения контрольной переменной функционала, как-то связанного с дисперсией оценки градиента, не ново: некоторые идеи подобного рода возникали, например, в [43]. Несмотря на это, практическая имплементация и теоретические исследования этого подхода на текущий момент отсутствуют в литературе.

1.4.1 Основные теоретические результаты

Главный объект нашего исследования – использование эмпирической дисперсии вместо критерия A2C. Говоря о эмпирической дисперсии, мы можем предложить два функционала для обучения бейзлайна:

$$V_{n,K}^{EVv}(\theta,\phi) := \frac{1}{K}\sum_{k=1}^{K}\left\|\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2 - \frac{1}{K^2}\left\|\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2, \tag{30}$$

$$V_{K}^{EVm}(\theta,\phi) := \frac{1}{K}\sum_{k=1}^{K}\left\|\widetilde{\nabla}^{b_\phi}J(X_n^{(k)})|_\theta\right\|_2^2. \tag{31}$$

Оба функционала, как может быть показано, являются несмещёнными оценками настоящей дисперсии оценки градиента, определяемой для случайного вектора Y

$$V(Y) := \mathbb{E}\left[\|Y - \mathbb{E}[Y]\|_2^2\right].$$

Соответствующие градиентные алгоритмы могу быть описаны как

$$\theta_{n+1} = \theta_n + \alpha_n\frac{1}{K}\sum_{k=1}^{K}\widetilde{\nabla}^{b_\phi}J(X_n^{(k)}), \tag{32}$$

$$\phi_{n+1} = \phi_n - \beta_n\nabla_\phi V_K^{EV}(\phi,\theta)|_{\phi_n,\theta_n}. \tag{33}$$

Мы получили два метода. Первый использует полную дисперсию V_K^{EVv} и называется EVv, второй называется EVm и использует V_K^{EVm}, тот же функционал, но без второго члена. Отметим, что алгоритм с критерием EVv работоспособен только если $K \geq 2$, в противном случае мы пытаемся оценить дисперсию по одному наблюдению. Можно отметить несколько простых фактов. Во-первых, оказывается, что при некоторых технических предположениях функционал A2C является верхней оценкой (с точностью до константы) функционалов EV (Утв.5 в [34]). Во-вторых, можно показать,

что если схема сходится к локальному оптимуму, то EVm и EVv асимптотически(при $n \to \infty$) эквивалентны, поскольку второй член дисперсии – это квадрат нормы настоящего градиента, который сходится к нулю 0.

Главный теоретический результат – оценка с высокой вероятностью для разности дисперсии оценки градиента с приближённо построенной контрольной переменной и лучшей в классе для шага n. Прежде всего, упростим нотацию. Будем далее обозначать оценку градиента как $h : \mathbb{R}^d \to \mathbb{R}^D$, класс оценок (полученных при разных бейзлайнах) как $\mathcal{H}$ и положим $\mathcal{E} = \mathbb{E}[h(X)] = \nabla_\theta J$, поскольку оценка градиента должна быть несмещённой. Чтобы снизить дисперсию оценки градиента, мы бы хотели выбирать на шаге n самую лучшую оценку в классе

$$h^* = \arg\min_{h \in \mathcal{H}} V(h),$$

где дисперсия V определена для каждого $h \in \mathcal{H}$ как

$$V(h) := \mathbb{E}\left[\|h(X) - \mathcal{E}\|^2\right],$$

со случайным вектором X, составленным из состояний, действий и наград, описанным ранее. Чтобы попробовать решить приведённую оптимизационную задачу, мы используем эмпирический аналог дисперсии и определяем

$$\hat{h} := \arg\min_{h \in \mathcal{H}} V_K(h)$$

с эмпирической дисперсией, определённой как

$$V_K(h) := \frac{1}{K-1} \sum_{k=1}^{K} \|h(X^{(k)}) - P_K h\|^2$$

с P_K в качестве эмпирической меры, так что мы можем обозначать выборочное среднее более кратко как

$$P_K h := \frac{1}{K} \sum_{k=1}^{K} h(X^{(k)}).$$

Приведём несколько ключевых предположений.

A 6. Класс $\mathcal{H}$ состоит из ограниченных функций:

$$\sup_{x \in \mathcal{X}} \|h(x)\| \le b, \quad \forall h \in \mathcal{H}.$$

A 7. Решение h_* единственно и $\mathcal{H}$ имеет звёздную форму(star-shaped) вокруг h_*:

$$\alpha h + (1-\alpha) h_* \in \mathcal{H}, \quad \forall h \in \mathcal{H},\ \alpha \in [0,1].$$

A 8. Класс $\mathcal{H}$ имеет покрытие полиномиального размера: существуют $\alpha \ge 2$ и $c > 0$ такие, что для всех $u \in (0, b]$,

$$\mathcal{N}(\mathcal{H}, \|\cdot\|_{L^2(P_K)}, u) \le \left(\frac{c}{u}\right)^\alpha \ a.s.,$$

где

$$\|h\|_{L^2(P_K)} = \sqrt{P_K \|h\|_2^2}$$

Теорема 8. *В предположениях 6-8 с вероятностью как минимум* $1 - 4e^{-t}$ *верно*

$$V(h_K) - V(h_*) \leq \max_{j=1,\dots,4} \beta_j(t)$$

с

$$\beta_1 \leq C_1 \frac{\log K}{K},\ \beta_2 \leq C_2 \frac{\log K}{K},$$
$$\beta_3(t) = \frac{C_3(t+1)}{3K}, \beta_4(t) = \frac{C_4 t}{K},$$

где константы C_1, C_2, C_3, C_4 *не зависят от размерности* D *или размера выборки* K *и определены в статье.*

Это позволяет утверждать, что при росте K дисперсия сходится к дисперсии h_*. С точки зрения практики, Теорема 8 прежде всего даёт некоторую гарантию надёжности алгоритма. Во-вторых, из неё следует, что при достаточно большом K можно добиться ещё большего снижения дисперсии.

1.4.2 Численные эксперименты

Мы исследуем работу EV-алгоритмов на нескольких классических образцовых задачах:

- Gym Minigrid [16] (`Unlock-v0, GoToDoor-5x5-v0`);
- Gym Classic Control [14] (`CartPole-v1, LunarLander-v2, Acrobot-v1`).

Для каждой из них мы предоставляем графики средних суммарных наград, иллюстрирующие процесс обучения, исследование градиентной дисперсии дисперсии суммы наград, а также данные о времени работы. Здесь для краткости мы предоставляем только самые важные результаты, но в Приложениях к диссертации можно найти больше экспериментов, а также детали имплементации. Программный код и все конфигурации можно найти на GitHub [30].

Обзор. Мы рассматриваем несколько ключевых индикаторов работы алгоритма.

1. **Средняя сумма наград.** Они вычисляются на каждой эпохе n, основываясь на наградах, полученных в течении обучения и усреднённых по 40 независимым запускам обучения. Средняя сумма наград показывает, как хорошо алгоритм взаимодействует со средой и собирает награды.
2. **Стандартное отклонение суммы наград.** Вычисляются точно так же, но вместо среднего по 40 запускам считается выборочное стандартное отклонение. данное значение показывает как устойчиво проходит обучение: высокие значения показывают, что бывают достаточно часто резкие взлёты или падения наград.

3. **Дисперсия градиента** замеряется каждые 200 эпох используя оценку (31), вычисленную на отдельном множестве из 50 траекторий, полученных с релевантной политикой. Это ключевой индикатор в обсуждениях снижения дисперсии. Насколько мы можем судить, наша работа является первой в RL-сообществе, представляющая подобные результаты для классических образцовых задач. Результирующие кривые усреднены по 40 независимым запускам обучения алгоритма.

4. **Степень снижения дисперсии.** Вместе с оценённой дисперсией градиента мы также вычисляем степень снижения дисперсии как выборочную дисперсию оценки градиента с бейзлайном, делённую на выборочную дисперсию оценки градиента без бейзлайна (то есть, полагая $b_\phi = 0$) при вычислениях дисперсии градиента. Степень снижения – главный индикатор в исследованиях по снижению дисперсии в Монте-Карло и MCMC.

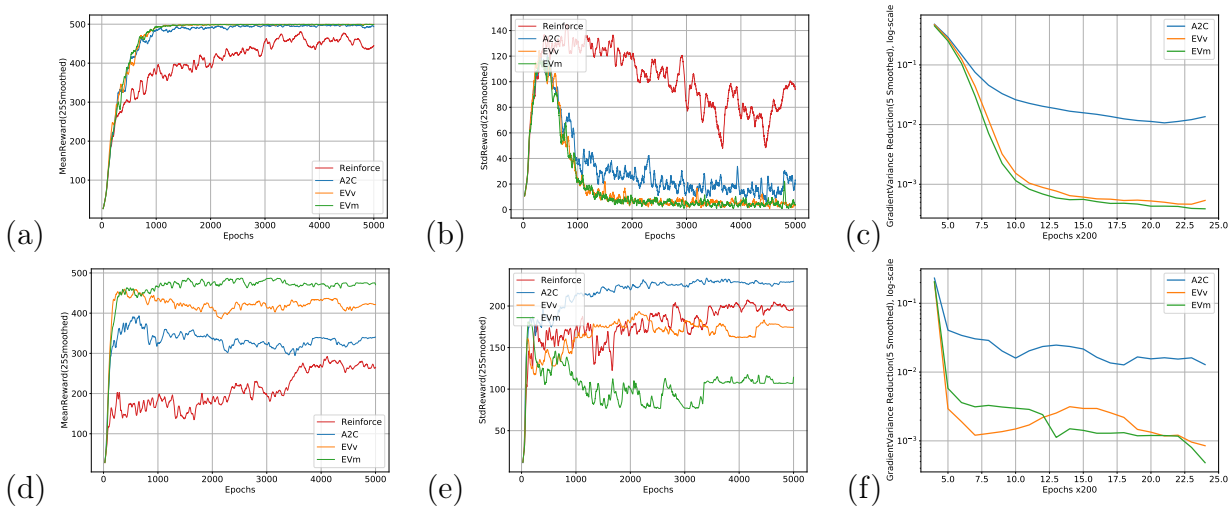

Рис. 3: Графики для среды `CartPole`: (a,b,c) показывают средние суммы наград, стандартное отклонение наград и степень снижения дисперсии градиента для конфигурации config5; (d,e,f) показывают то же для конфигурации config8.

Работа алгоритма. Наблюдая за средними суммами наград, можно заметить, что EV-алгоритмы как минимум так же хороши, как и A2C. В среде `CartPole` (Fig. 3) мы провели несколько экспериментов и представляем здесь две самых иллюстративных конфигурации политики: одну с более простой моделью политики (config5, см Рис. 3(a,b,c)) и другую с более сложной моделью (config8, see Fig. 3(d,e,f)). В первом случае A2C и EV работают с очень похожим успехом, но во втором случае агент с EV-методом снижения дисперсии учится существенно быстрее и мы получаем улучшение примерно на 50% по сравнению с A2C и 75% по сравнению с простейшим агентом Reinforce в конце и ещё больше в течении обучения. Этот феномен лучшей работы EV в среде `CartPole` при более сложной модели политики наблюдается очень часто, детальнее можно посмотреть в Приложениях. Что касается среды `Acrobot` (см. Рис. 4(a)), мы видим, что EV-алгоритмы дают лучший прирост скорости обучения. В начале EVm позволяет учиться быстрее, но в конце агент так же успешен, как и A2C. Одна из причин такого поведения может состоять в том, что размер шага

оптимизации (learning rate) становится слишком малым и агент достигает потолка в обучении. Среда `Unlock` (Рис. 5(a)) – это пример среды, где все алгоритмы работают примерно одинаково: в терминах награды мы не видим значительных улучшений по сравнению с Reinforce.

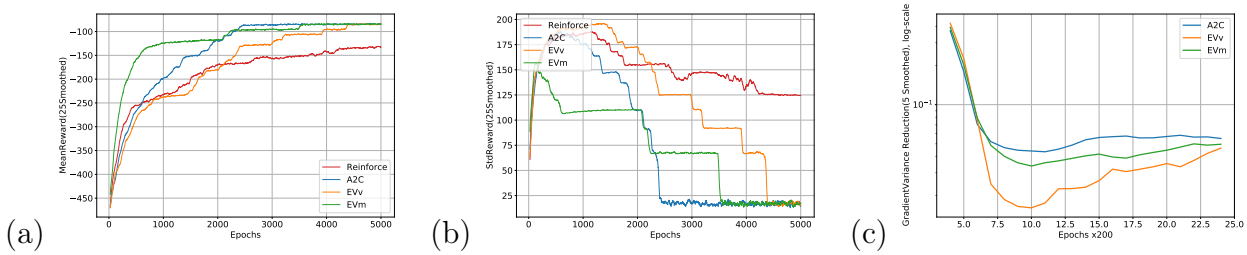

Рис. 4: Графики для среды `Acrobot`: (a) отображает средние суммы наград, (b) показывает стандартное отклонение наград и (c) демонстрирует степень снижения дисперсии градиента.

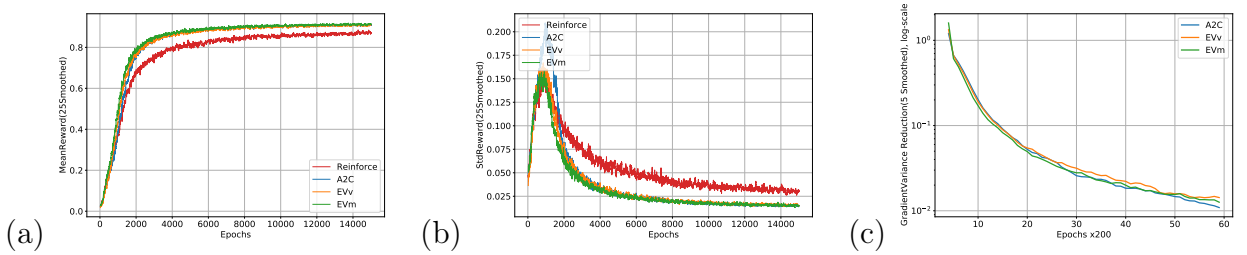

Рис. 5: Графики для среды `Unlock`: (a) отображает средние суммы наград, (b) показывает стандартное отклонение наград и (c) демонстрирует степень снижения дисперсии градиента.

Стабильность обучения. Согласно полученным данным по стандартному отклонению наград (Рис. 3(b,e),4(b),5(b)), EV-методы лучше себя показывают с точки зрения устойчивости обучения:резкие провалы случаются реже, чем в A2C. Это хорошо видно в среде `CartPole` на Рис. 3(b,e), где стандартное отклонение примерно в два раза ниже, чем в A2C. Это верно для обеих конфигураций. Рис. 4, иллюстрирующий эксперименты со средой `Acrobot`, показывает, что до достижения потолка EVm всё ещё имеет дисперсию ниже, чем в A2C или Reinforce. В среде `Unlock`, представленной на Рис. 5(b) значительной разницы.

Дисперсия градиента и её влияние. Первое, что мы можем заметить, смотря на графики дисперсии градиента, – это то, что A2C и EV одинаково хорошо уменьшают дисперсию в среде `Unlock`. В среде `CartPole` (см. Рис. 3(c,f)), однако, мы видим, что результаты EV и A2C сильно отличаются: EV уменьшает дисперсию в 100-1000 раз в обеих конфигурациях. Похожий результат мы наблюдаем во всех экспериментах со средой `CartPole`. Можно заметить, что в `Unlock`, показанной на Рис. 5 дисперсия также уменьшается в 10-100 раз, но это очень слабо отражается на прогрессе обучения. Этот факт демонстрирует, что в некоторых средах обучение не восприимчиво к снижению дисперсии; кроме того, возможно, что дисперсия снижается недостаточно сильно. Наконец, мы хотели бы отметить, что дисперсия наград,

измеренная в предыдущем параграфе, не является надёжным индикатором снижения дисперсии. Дисперсия наград заметно уменьшается по отношению к Reinforce только в среде `CartPole`, в других мы этого не наблюдаем. Следовательно, дисперсию наград нельзя использовать для исследования снижения дисперсии в RL. Тем не менее, дисперсия наград и дисперсия градиента связаны и является открытым вопросом, как именно.

Выводы

Что касается первой из поставленных целей диссертационного исследования, для задач оптимальной остановки в дискретном времени нам удалось установить ST-решаемость(semitractability) с помощью алгоритма WSM при слабых предположениях на Марковскую цепь: требуется лишь наличие плотности. В частном случае, когда функция продолжения (вложенное условное матожидание) является является бесконечно гладкой функцией, многие регрессионные алгоритмы, включая метод Лонгштаффа-Шварца, также дают ST-решаемость задачи оптимальной остановки в дискретном времени. Несмотря на это, как было показано, их недостаточно для задачи оптимальной остановки в непрерывном времени – регрессионные алгоритмы дают бесконечный ST-индекс(semitractability index), в то время как индекс метода WSM остаётся ограниченным. Эффект этого можно чётко видеть на экспериментах.

Во втором направлении мы получили улучшенные неасимптотические оценки сходимости линейной двумасштабной схемы стохастической аппроксимации в случае мартингального и марковского шума при более слабых, чем в литературе предположениях. Наш анализ через вывод и оценку рекурсий является очень точным и позволяет получить лучшие скорости сходимости.

Что касается третьей цели, мы предложили использовать эмпирическую дисперсию и использовать EV-методы. Мотивация использования EV-алгоритмов больше о действительном снижении дисперсии, нежели в случае A2C, и по всем критериям они показывают по крайней мере такие же хорошие результаты, как A2C в терминах получаемых наград и снижения дисперсии. Для EV-методов впервые в литературе мы предложили вероятностные оценки для дисперсии оценки градиента при достаточно слабых предположениях. EV-алгоритмы показывают себя более стабильными в процессе обучения, что позволяет снизить возможность внезапных провалов. Мы также впервые представили эмпирическое исследование снижения дисперсии оценки градиента в классических образцовых задачах. Наши результаты показывают, что снижение дисперсии может помочь в обучении, но иногда некоторые специфичные для среды черты не позволяют получить с его помощью ощутимого эффекта в скорости обучения. Следовательно, техники снижения дисперсии должны использоваться, но не всегда и точные предпосылки для этого пока неизвестны.

Список литературы

[1] Ankush Agarwal and Sandeep Juneja. Comparing optimal convergence rate of stochastic mesh and least squares method for bermudan option pricing. In *Proceedings of the 2013 Winter Simulation Conference: Simulation: Making Decisions in a Complex World*, WSC '13, page 701–712. IEEE Press, 2013.

[2] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[3] Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 30–37, 1995.

[4] Vlad Bally, Gilles Pagès, and Jacques Printems. A quantization tree method for pricing and hedging multidimensional american options. *Mathematical Finance*, 15(1):119–168, 2005.

[5] Denis Belomestny, Leonid Iosipoi, Quentin Paris, and Nikita Zhivotovskiy. Empirical variance minimization with applications in variance reduction and optimal control. *Bernoulli*, 28(2):1382 – 1407, 2022.

[6] Denis Belomestny, Maxim Kaledin, and John Schoenmakers. Semitractability of optimal stopping problems via a weighted stochastic mesh algorithm. *Mathematical Finance*, 30(4):1591–1616, 2020.

[7] Michel Benaïm. Dynamics of stochastic approximation algorithms. *Séminaire de probabilités de Strasbourg*, 33:1–68, 1999.

[8] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. 12 2019.

[9] D. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific optimization and computation series. Athena Scientific, 2019.

[10] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692, 2018.

[11] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[12] Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[13] Mark Broadie and Paul Glasserman. A stochastic mesh method for pricing high-dimensional american options. *Journal of Computational Finance*, 7:35–72, 2004.

[14] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[15] Ching-An Cheng, Xinyan Yan, and Byron Boots. Trajectory-wise control variates for variance reduction in policy gradient methods. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1379–1394. PMLR, 30 Oct–01 Nov 2020.

[16] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. `https://github.com/maximecb/gym-minigrid`, 2018.

[17] Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research*, 21(52):1–51, 2020.

[18] E. Clément, D. Lamberton, and Philip Protter. An analysis of a least squares regression algorithm for american option pricing. *Finance and Stochastics*, 17, 01 2002.

[19] Gal Dalal, Balazs Szorenyi, and Gugan Thoppe. A tale of two-timescale reinforcement learning with the tightest finite-time bound. *arXiv preprint arXiv:1911.09157*, 2019.

[20] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233, 2018.

[22] Thinh T Doan. Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation. *arXiv preprint arXiv:1912.10583*, 2019.

[23] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

[24] Daniel Egloff, Michael Kohler, and Nebojsa Todorovic. A dynamic look-ahead monte carlo algorithm for pricing bermudan options. *The Annals of Applied Probability*, 17(4):1138–1171, 2007.

[25] Yannis Flet-Berliac, reda ouhamma, odalric-ambrym maillard, and Philippe Preux. Learning value functions in deep policy gradients using residual variance. In *International Conference on Learning Representations*, 2021.

[26] Vincent Francois, David Taralla, Damien Ernst, and Raphael Fonteneau. Deep reinforcement learning solutions for energy microgrids management. 12 2016.

[27] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354, 2018.

[28] Paul Glasserman. *Monte Carlo methods in financial engineering*. Springer, New York, 2004.

[29] David A. Goldberg and Yilun Chen. Beating the curse of dimensionality in options pricing and optimal stopping, 2018.

[30] Alexander Golubev and Maksim Kaledin. EVRLLib, a library implementing policy gradient algorithms in Reinforcement Learning. `https://github.com/DJAlexJ/EVRLLib`, June 2022.

[31] Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *J. Mach. Learn. Res.*, 5:1471–1530, December 2004.

[32] Shixiang Gu, Timothy P. Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-prop: Sample-efficient policy gradient with an off-policy critic. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[33] Harsh Gupta, R Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4706–4715, 2019.

[34] Maxim Kaledin, Alexander Golubev, and Denis Belomestny. Variance reduction for policy-gradient methods via empirical variance minimization. *arXiv:2206.06827v2*, 2022.

[35] Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2144–2203. PMLR, 09–12 Jul 2020.

[36] Beom Kim, Yong-Ki Ma, and Hi Choe. A simple numerical method for pricing an american put option. *Journal of Applied Mathematics*, 2013, 01 2013.

[37] Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1008–1014. MIT Press, 2000.

[38] Vijay R. Konda and John N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *Ann. Appl. Probab.*, 14(2):796–819, 05 2004.

[39] Harold Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

[40] Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.

[41] Tanguy Levent, Philippe Preux, Erwan Le Pennec, Jordi Badosa, Gonzague Henri, and Yvan Bonnassieux. Energy Management for Microgrids: a Reinforcement Learning Approach. In *ISGT-Europe 2019 - IEEE PES Innovative Smart Grid Technologies Europe*, pages 1–5, Bucharest, France, September 2019. IEEE.

[42] Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient TD algorithms. In *UAI*, pages 504–513, 2015.

[43] Hao Liu, Yihao Feng, Yi Mao, Dengyong Zhou, Jian Peng, and Qiang Liu. Action-dependent control variates for policy optimization via stein identity. In *International Conference on Learning Representations*, 2018.

[44] Francis Longstaff and Eduardo Schwartz. Valuing american options by simulation: A simple least-squares approach. *Review of Financial Studies*, 14:113–47, 02 2001.

[45] Hongzi Mao, Shaileshh Bojja Venkatakrishnan, Malte Schwarzkopf, and Mohammad Alizadeh. Variance reduction for reinforcement learning in input-driven environments. In *International Conference on Learning Representations*, 2019.

[46] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, 2016.

[47] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[48] Abdelkader Mokkadem, Mariane Pelletier, et al. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3):1671–1702, 2006.

[49] Erich Novak and Henryk Wo'zniakowski. *Tractability of multivariate problems. Vol. 1: Linear information.* 01 2008.

[50] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4026–4035, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[51] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[52] John Rust. Using randomization to break the curse of dimensionality. *Econometrica*, 65(3):487–516, 1997.

[53] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

[54] Shijing Si, Chris. J. Oates, Andrew B. Duncan, Lawrence Carin, and François-Xavier Briol. Scalable control variates for monte carlo methods via stochastic optimization, 2021.

[55] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 10 2017.

[56] R. Srikant and Lei Ying. Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning. In *Conference on Learning Theory*, 2019.

[57] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[58] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018.

[59] Richard S Sutton, Hamid Maei, and Csaba Szepesvári. A convergent o(n) temporal-difference algorithm for off-policy learning with linear function approximation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008.

[60] Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 993–1000, New York, NY, USA, 2009. Association for Computing Machinery.

[61] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000.

[62] Vladislav Tadic. Almost sure convergence of two time-scale stochastic approximation algorithms. *Proceedings of the 2004 American Control Conference*, 4:3802–3807 vol.4, 2004.

[63] Vladislav Tadic. Asymptotic analysis of temporal-difference learning algorithms with constant step-sizes. *Machine Learning*, 63:107–133, 05 2006.

[64] Nizar Touzi. Optimal stochastic control, stochastic target problems, and backward sde. *Fields Institute Monographs*, 29, 01 2013.

[65] Lloyd Trefethen. Multivariate polynomial approximation in the hypercube. *Proceedings of the American Mathematical Society*, 145, 08 2016.

[66] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, May 1997.

[67] John Tsitsiklis and Benjamin Roy. Regression methods for pricing complex american style options. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 12:694–703, 02 2001.

[68] George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5015–5024. PMLR, 10–15 Jul 2018.

[69] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, Nov 2019.

[70] L. Weaver and N. Tao. The optimal reward baseline for gradient·based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2001.

[71] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, May 1992.

[72] Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In *International Conference on Learning Representations*, 2018.

[73] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 541–551, Tel Aviv, Israel, 22–25 Jul 2020. PMLR.

[74] Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*, pages 10633–10643, 2019.

[75] Daniel Zanger. Quantitative error estimates for a least-squares monte carlo algorithm for american option pricing. *Finance and Stochastics*, 17, 07 2013.