КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 2 С. 273–288
DOI: 10.20537/2076-7633-2026-18-2-273-288

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 681.5.015

Применение алгоритма QUBO для отбора траекторий обучения с подкреплением методом Монте-Карло

Я. А. Холодов[1,a], Х. Саллум[2,b], А. Джнади[2,c], К. Ю. Хубиев[1,d], А. Петренко[3,e]

[1]Научно-технологический университет «Сириус»,
Россия, 354340, Краснодарский край, Федеральная территория «Сириус», пгт Сириус, проспект Олимпийский, д. 1
[2]Университет Иннополис,
Россия, 420500, Республика Татарстан, Верхнеуслонский муниципальный район, город Иннополис, ул. Университетская, д. 1
[3]Санкт-Петербургский федеральный исследовательский центр Российской академии наук,
Россия, 199178, г. Санкт-Петербург, 14-я линия В. О., д. 39

E-mail: [a] kholodov.ya@talantiuspeh.ru, [b] h.salloum@innopolis.university, [c] a.jnadi@innopolis.university, [d] hubiev.k@talantiuspeh.ru, [e] a.petrenko1999@rambler.ru

Получено 18.12.2025, после доработки — 06.04.2026.
Принято к публикации 08.04.2026.

Метод Монте-Карло (Monte Carlo, MC) в обучении с подкреплением показывает низкую эффективность при высокой сложности обучающей выборки — в средах с редким вознаграждением, большим пространством состояний и коррелирующими траекториями. Эти ограничения приводят к повышенной вариативности оценок возврата и существенно замедляют процесс сходимости, особенно в задачах, где требуется выделить наиболее информативные эпизоды из большого множества доступных данных. При прямом использовании всех траекторий возникает избыток информации, что ухудшает качество итоговых оценок и увеличивает вычислительную нагрузку. В данной работе мы предлагаем подход, позволяющий преодолеть указанные проблемы за счет оптимизации отбора обучающих данных и структурирования выборки перед применением классического метода Монте-Карло. Задача отбора обучающих траекторий формулируется как квадратичная неограниченная бинарная оптимизация (*Quadratic Unconstrained Binary Optimization*, QUBO) и решается с помощью алгоритма квантового отжига. Предлагаемый метод MC + QUBO интегрирует комбинаторный фильтрующий шаг в стандартную процедуру оценки: из множества потенциальных траекторий выбирается поднабор, максимизирующий суммарное вознаграждение, обеспечивая при этом достаточное покрытие пространства состояний и снижение взаимной корреляции эпизодов. В QUBO-формулировке линейные члены поощряют включение эпизодов с высоким значением возврата, тогда как квадратичные члены регулируют разнообразие и баланс траекторий, уменьшая риск переобучения на узком подмножестве данных. В качестве решателей из категории «черного ящика» используются алгоритмы симуляции квантового отжига (*Simulated Quantum Annealing*, SQA) и симулированная бифуркация (*Simulated Bifurcation*, SB), что позволяет эффективно решать задачи с большим числом потенциальных эпизодов и быстро находить приближенные оптимальные решения. Эксперименты в среде *GridWorld* показывают, что MC + QUBO превосходит классический метод Монте-Карло по скорости сходимости, устойчивости оценок и качеству итогового обучения, демонстрируя потенциал квантовой оптимизации как инструмента повышения эффективности принятия решений в задачах обучения с подкреплением.

Ключевые слова: метод Монте-Карло, квантовый отжиг, квантовые вычисления, обучение с подкреплением, QUBO

Работа выполнена при финансовой поддержке проекта «Технологии противодействия ранее неизвестным квантовым киберугрозам», реализуемого в рамках государственной программы федеральной территории «Сириус» «Научно-технологическое развитие федеральной территории “Сириус”» (соглашение № 23-03 от 27.09.2024 г.).

COMPUTER RESEARCH AND MODELING
2026 VOL. 18 NO. 2 P. 273–288
DOI: 10.20537/2076-7633-2026-18-2-273-288

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 681.5.015

Quantum-inspired episode selection for Monte Carlo reinforcement learning via QUBO optimization

Ya. A. Kholodov[1,a], H. Salloum[2,b], A. Jnadi[2,c], K. Yu. Khubiev[1,d], A. Petrenko[3,e]

[1]Sirius University of Science and Technology,
1 Olimpiyskiy ave., Sirius, Sirius Federal Territory, Krasnodar region, 354340, Russia
[2]Innopolis University,
1 Universitetskaya st., Innopolis, Tatarstan, 420500, Russia
[3]St. Petersburg Federal Research Center of the Russian Academy of Science,
39 14th Line V. O., St. Petersburg, 199178, Russia

E-mail: [a] kholodov.ya@talantiuspeh.ru, [b] h.salloum@innopolis.university, [c] a.jnadi@innopolis.university, [d] hubiev.k@talantiuspeh.ru, [e] a.petrenko1999@rambler.ru

Received 18.12.2025, after completion — 06.04.2026.
Accepted for publication 08.04.2026.

Monte Carlo (MC) reinforcement learning suffers from high sample complexity, especially in environments with sparse rewards, large state spaces, and strongly correlated trajectories that reduce the statistical efficiency of return estimation. These well-known limitations often lead to slow convergence and unstable learning dynamics, particularly in settings where only a small fraction of collected trajectories is actually informative for policy improvement. A key challenge is therefore to identify a compact yet diverse subset of episodes that contributes most to the accuracy of value estimates while preserving sufficient exploration of the environment. To address this challenge, we reformulate episode selection as a Quadratic Unconstrained Binary Optimization (QUBO) problem and solve it using quantum-inspired sampling techniques. Our method, MC + QUBO, inserts a combinatorial filtering step into the standard MC policy-evaluation pipeline: given a batch of trajectories, it selects a subset that maximizes cumulative reward and encourages broad state-space coverage. This selection procedure is expressed as a QUBO model, where linear terms favor high-return episodes, quadratic terms penalize redundancy between trajectories, and additional coupling terms can be used to enforce coverage-related constraints or promote structural diversity. Within this framework, we investigate two black-box QUBO solvers: Simulated Quantum Annealing (SQA), which emulates tunneling-based exploration of the search landscape, and Simulated Bifurcation (SB), a dynamical-systems-based iterative optimization method. Both solvers demonstrate the ability to efficiently navigate the combinatorial structure of the trajectory-selection problem and to handle batch sizes that are otherwise computationally expensive for exhaustive or deterministic search. Experiments in a finite-horizon GridWorld environment show that MC + QUBO consistently outperforms vanilla MC in convergence speed, stability of return estimates, and final policy quality. These results highlight the promise of quantum-inspired optimization as a practical decision-making subroutine within reinforcement-learning algorithms, offering a scalable way to improve sample efficiency without modifying the underlying learning paradigm.

Keywords: method Monte Carlo, quantum annealing, quantum computation, reinforcement learning, QUBO

Citation: *Computer Research and Modeling*, 2026, vol. 18, no. 2, pp. 273–288 (Russian).

This work was obtained with the financial support of the project “Technologies for countering previously unknown quantum cyber threats”, implemented within the framework of the state program of the “Sirius” Federal Territory “Scientific and technological development of the ‘Sirius’ Federal Territory” (Agreement No. 23-03 dated September 27, 2024).

Введение

Обучение с подкреплением (Reinforcement Learning, RL) предоставляет формальный аппарат для решения задач принятия решений в стохастических средах, моделируемых марковскими процессами (Markov Decision Process, MDP) [Puterman, 1994; Bertsekas, Tsitsiklis, 1996; Singh, Sutton, 1996; Sutton, Barto, 1998]. В рамках этого подхода методы оценки функций ценности на основе метода Монте-Карло (Monte Carlo, MC) занимают особое место благодаря своей концептуальной простоте и несмещенности оценок возвратов [Bertsekas, Tsitsiklis, 1996; Sutton, Barto, 1998]. Однако практическое применение MC-методов сталкивается с фундаментальной проблемой: при наличии большого числа доступных эпизодов использование всей выборки оказывается не только вычислительно затратным, но и статистически неэффективным.

Ключевая причина этой неэффективности заключается в том, что вклад отдельных траекторий в точность оценки функции ценности не является аддитивным. Эпизоды, порожденные одной и той же политикой, часто сильно коррелированы, посещают близкие подмножества состояний и несут избыточную информацию [Kearns, Singh, 2000; Greensmith et al., 2004; Liu, 2021]. В результате стандартное усреднение по всей выборке приводит к росту дисперсии оценок и замедлению сходимости, особенно в задачах с редкими вознаграждениями и большим пространством состояний [Thomas, Brunskill, 2016; Jiang, Li, 2016; Winnicki, Srikant, 2023]. Таким образом, возникает естественный вопрос: какое подмножество эпизодов действительно следует использовать для обновления политики?

С формальной точки зрения задача отбора эпизодов не сводится к независимой фильтрации «плохих» траекторий. Выбор одного эпизода изменяет ценность других, поскольку их информативность определяется перекрытием посещаемых состояний и корреляцией возвратов. Это означает, что оптимальный отбор эпизодов представляет собой комбинаторную задачу с попарными взаимодействиями, близкую по структуре к задачам выбора признаков и максимального покрытия, которые являются NP-трудными [Garey, Johnson, 1979; Feige, 1998; Guyon, Elisseeff, 2003; Das, Kempe, 2011; Krause, Golovin, 2014]. Следовательно, эвристические критерии локальной важности или жадные алгоритмы не могут гарантировать получение репрезентативного подмножества при росте размера выборки.

В данной работе мы предлагаем формализовать задачу отбора эпизодов в методе Монте-Карло как задачу квадратичной неограниченной бинарной оптимизации (Quadratic Unconstrained Binary Optimization, QUBO). В этой формулировке каждому эпизоду τ_i сопоставляется бинарная переменная $x_i \in \{0, 1\}$, указывающая, включен ли данный эпизод в обучающую подвыборку. Целевая функция QUBO строится таким образом, чтобы одновременно учитывать индивидуальную информативность эпизодов и их попарную избыточность.

Линейные члены QUBO-модели отвечают за вклад отдельных эпизодов и кодируют их «ценность» с точки зрения обучения — например, суммарное вознаграждение или иной прокси-показатель полезности. Квадратичные члены описывают взаимодействия между эпизодами и штрафуют за совместный выбор траекторий, обладающих высокой степенью сходства, например по множеству посещенных состояний. Таким образом, минимизация QUBO-функции реализует явный компромисс между качеством и разнообразием выборки, что принципиально невозможно при простом усреднении по всем эпизодам.

Преимущество QUBO-формализации заключается в ее универсальности и вычислительной выразительности. Широкий класс NP-трудных задач комбинаторной оптимизации допускает естественное представление в виде QUBO или эквивалентной модели Изинга [Kadowaki, Nishimori, 1998; Morita, Nishimori, 2008; Lucas, 2014]. Это в свою очередь открывает возможность применения специализированных алгоритмов оптимизации, включая алгоритмы квантового отжига и квантово-вдохновленные методы, такие как симуляция квантового отжига (Simulated

Quantum Annealing, SQA) и симуляция бифуркации (Simulated Bifurcation, SB) [Kochenberger et al., 2014; McGeoch, 2014; Goto et al., 2019; Hauke et al., 2020; Zlokapa, Carleo, 2021; Glover, Kochenberger, 2022; Salloum et al., 2025].

Важно подчеркнуть, что в предлагаемом подходе QUBO не используется как внешний эвристический фильтр, а интегрируется непосредственно в контур обучения с подкреплением. На каждом шаге алгоритма генерируется набор эпизодов, после чего формируется QUBO-задача, описывающая оптимальный выбор подмножества этих эпизодов. Решение QUBO-задачи определяет, какие траектории используются для обновления функции ценности методом Монте-Карло, после чего политика обновляется стандартным образом. Такой механизм превращает этап выборки данных из пассивного накопления информации в активную процедуру комбинаторной оптимизации.

Следует отметить, что, несмотря на наличие работ, посвященных применению QUBO и квантовой оптимизации в задачах управления и планирования [Kochenberger et al., 2014; McGeoch, 2014; Crosson, Harrow, 2016; Goto et al., 2019; Glover, Kochenberger, 2022], их использование для структурирования обучающих данных в RL до настоящего времени рассматривалось ограниченно. В существующих подходах отсутствует явная связь между параметрами QUBO-модели и статистическими свойствами MC-оценок, такими как дисперсия и корреляция возвратов.

В данной работе мы восполняем этот пробел, демонстрируя, что QUBO-отбор эпизодов позволяет систематически уменьшать избыточность выборки, снижать влияние скоррелированных траекторий и ускорять сходимость метода Монте-Карло без изменения базовой схемы обучения. Тем самым обучение с подкреплением переосмысливается как задача, в которой ключевую роль играет не только стратегия агента, но и оптимизационно управляемый процесс отбора данных.

Модель Изинга [Cipra, 1987], происходящая из статистической механики, представляет собой фундаментальную основу для задач комбинаторной оптимизации [Bashar, Shukla, 2023]. Пусть модель определена на ненаправленном графе $G = (V, E)$ с числом вершин $|V| = n$. Тогда гамильтониан модели со спинами $s \in \{\pm 1\}^n$ имеет вид

$$H_{\text{Ising}}(s) = -\sum_{(i,\, j)\in E} J_{ij} s_i s_j - \sum_{i\in V} h_i s_i, \tag{1}$$

где $J_{ij} \in \mathbb{R}$ обозначает попарные взаимодействия, а $h_i \in \mathbb{R}$ — локальные силовые поля. При $J_{ij} \geqslant 0$ связи являются ферромагнитными, то есть спины стремятся к выравниванию.

В формулировке квадратичной неограниченной бинарной оптимизации (*QUBO*) гамильтониан принимает эквивалентный вид:

$$H_{\text{QUBO}}(\mathbf{x}) = \mathbf{x}^{\top} Q \mathbf{x} + \mathbf{q}^{\top} \mathbf{x}, \quad \mathbf{x} \in \{0,\, 1\}^n, \tag{2}$$

где $Q \in \mathbb{R}^{n\times n}$ — симметричная матрица, диагональные элементы которой могут задавать линейные смещения [Salloum et al., 2025].

Связь между спинами модели Изинга и бинарными переменными выражается следующим образом:

$$s_i = 2x_i - 1. \tag{3}$$

Подставляя выражение $s_i s_j = 4x_i x_j - 2x_i - 2x_j + 1$ в уравнение (1) и приводя подобные члены, получаем

$$H_{\text{Ising}}(\mathbf{s}) = \underbrace{\left(-\sum_{i<j} J_{ij} s_i s_j - \sum_i h_i s_i\right)}_{C} + \sum_i \left(2\sum_{j\neq i} J_{ij} - 2h_i\right) x_i + \sum_{i<j} (-4J_{ij}) x_i x_j, \tag{4}$$

где C — добавочная константа, не влияющая на процесс оптимизации. Таким образом, QUBO-модель записывается с обозначениями:

$$Q_{ij} = -4J_{ij} \quad (i < j), \quad q_i = 2\sum_{j \neq i} J_{ij} - 2h_i. \tag{5}$$

Замечание 1. Существуют различные соглашения о знаках для гамильтонианов моделей Изинга и QUBO; приведенные выше соотношения точны для выбранной здесь системы обозначений.

Рассмотрим задачу минимизации в виде

$$\text{MIN-QUBO:} \quad \min_{\mathbf{x} \in \{0,\,1\}^n} F(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{q}^\top \mathbf{x}. \tag{6}$$

Эта задача принадлежит к классу *NP*-трудных: ее точное решение за полиномиальное время означало бы, что $P = NP$. Связанная с ней задача принятия решения формулируется следующим образом:

$$\text{QUBO-Decision:} \quad \text{для данных } Q,\ \mathbf{q},\ T,\ \exists\, \mathbf{x} \in \{0,\,1\}^n \text{ такой, что } F(\mathbf{x}) \leqslant T? \tag{7}$$

Эта задача относится к классу *NP*-полных. Задача MIN-QUBO является *NP*-трудной, поскольку оптимизатор, решающий ее, должен соответствовать задаче принятия решения для любого входного набора.

Набросок доказательства (упрощение задачи разбиения). Для данных положительных целых чисел $a_1, \ldots, a_n$ и $K = \frac{1}{2}\sum_i a_i$ определим функцию

$$f(\mathbf{x}) = \left(\sum_{i=1}^{n} a_i x_i - K\right)^2 = \mathbf{x}^\top \left(\mathbf{a}\mathbf{a}^\top\right)\mathbf{x} - 2K\mathbf{a}^\top\mathbf{x} + K^2. \tag{8}$$

Положим $Q = \mathbf{a}\mathbf{a}^\top$ (ранг 1), $\mathbf{q} = -K\mathbf{a}$, константу $C = K^2$. Тогда $\min f(\mathbf{x}) = 0$ тогда и только тогда, когда сумма подвыборки равна K, что соответствует решению задачи разбиения.

Аппроксимация и особые случаи

- Некоторые подклассы (например, **MAX-CUT** с неотрицательными весами) допускают приближенные решения с постоянным коэффициентом точности (Goemans – Williamson, $\approx 0{,}878$) [Goemans, Williamson, 1995].
- Для произвольных знакопеременных весов гарантировать аппроксимацию, как правило, невозможно без введения структурных ограничений.
- Отдельные частные случаи (например, положительно полуопределенная матрица Q с разреженной структурой) могут быть решены эффективно.

Практические замечания. Поскольку точное решение при больших n зачастую невыполнимо, на практике применяются следующие подходы.

***Точные методы*:** целочисленное программирование, ветвление и границы — для задач малой и средней размерности.

***Релаксации*:** методы на основе полуположительно определенного программирования (*SDP*), спектральные методы — для получения нижних/верхних оценок с последующим округлением решений.

***Эвристики*:** физически мотивированные методы и метаэвристики, используемые для задач большой размерности с приближенным поиском лучших решений.

Мы рассматриваем два физических подхода к решению задач QUBO/Изинга: симуляцию квантового отжига (*Simulated Quantum Annealing*, SQA) [Crosson, Harrow, 2016] и симулированную бифуркацию (*Simulated Bifurcation*, SB) [Goto et al., 2019]. Цель данной работы состоит в интеграции подобных решателей в агентов обучения с подкреплением (*Reinforcement Learning*, RL), что позволит реализовать адаптивные стратегии, использующие преимущества физически обусловленной оптимизации в процессах принятия решений. Насколько нам известно, лишь ограниченное число исследований посвящено интеграции *RL* с квантовыми решателями, и существующие работы имеют ограничения как по охвату, так и по глубине анализа.

Мы начинаем с аппроксимации функции ценности в обучении с подкреплением и выбираем метод Монте-Карло в качестве отправной точки, поскольку он обеспечивает простой и несмещенный базовый результат, который далее может служить эталоном для качественного сравнения.

Симулятор SQA имитирует процесс квантового отжига (подробнее о концепте квантового отжига и его математическом формализме смотрите в [Kadowaki, Nishimori, 1998; Morita, Nishimori, 2008]) с помощью классических стохастических методов, таких как интеграл Монте-Карло вдоль траектории, а не на реальном квантовом компьютере.

Процесс обновления гамильтониана интерполируется между начальным гамильтонианом H_0 и целевым гамильтонианом H_P:

$$H_{\mathrm{SQA}}(t) = A(t)H_0 + B(t)H_P, \quad t \in [0, T], \tag{9}$$

с граничными условиями

$$A(0) = 1, \quad B(0) = 0, \quad A(T) = 0, \quad B(T) = 1. \tag{10}$$

Во время симуляции происходит выборка траекторий из функции эффективного квантового разбиения с использованием дискретной размерности «мнимого времени» для моделирования квантовых флуктуаций. Эти флуктуации позволяют симулировать эффект квантового туннелирования между локальными минимумами. Механизм туннелирования позволяет избегать областей квантового захвата (квантовых ловушек) по сравнению с энергетическими скачками через барьеры, которые реализуются за счет тепловой энергии в классических симуляциях метода квантового отжига.

Эффективность метода SQA зависит от профиля отжига ($A(t)$, $B(t)$), числа срезов Троттера (определяющих разрешающую способность квантовых флуктуаций) и стратегии обновления методом Монте-Карло. Важным является правильный баланс между исследованием (квантовые флуктуации с большой амплитудой) и использованием накопленного опыта (классическая сходимость) за t шагов. Несмотря на то что метод SQA не является квантовым в строгом смысле, он наследует многие свойства квантового отжига и является эффективным инструментом для программной реализации, независимой от конкретной вычислительной архитектуры.

С другой стороны, метод симулированных бифуркаций (SB) представляет собой классический подход для динамических систем, который переводит непрерывные переменные в дискретные состояния посредством явления бифуркации:

$$\dot{x}_i = y_i, \quad \dot{y}_i = -(\mu(t) - \lambda)x_i + \sum_j J_{ij}x_j - \gamma y_i. \tag{11}$$

Здесь параметр $\mu(t)$ плавно проходит через критическое значение $\mu_c = \lambda\rho(J)$ (спектральный радиус матрицы J), вызывая переход с нарушением симметрии и способствуя формированию

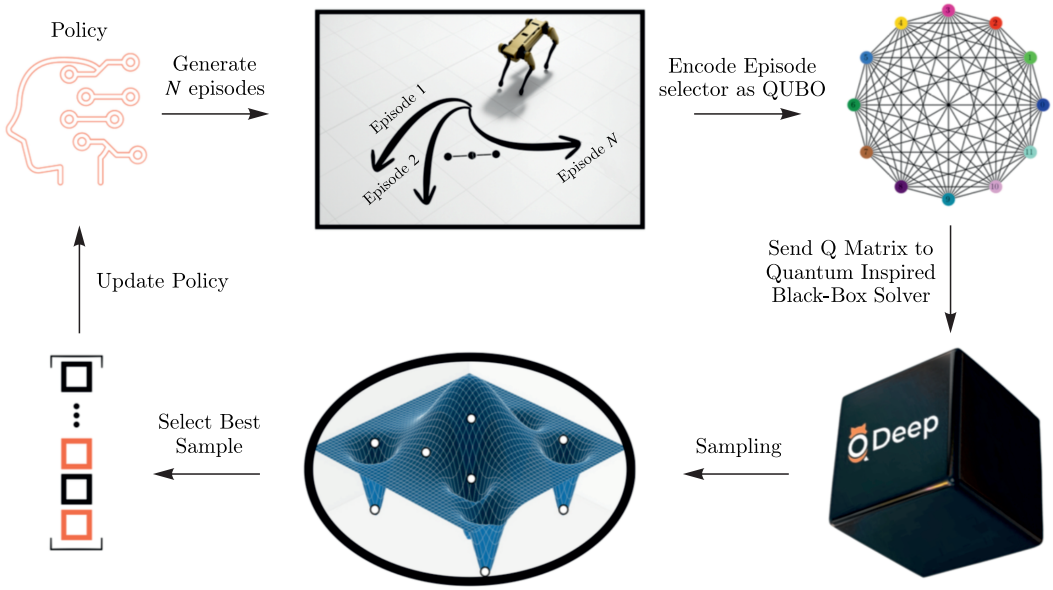

Рис. 1. Итеративное обучение с подкреплением с квантовым ассистированием: система генерирует несколько эпизодов, которые кодируются в виде задачи QUBO и затем решаются квантовым алгоритмом «черного ящика». Наилучший элемент выборки используется для обновления политики, после чего процесс повторяется итеративно с целью достижения оптимального результата

низкоэнергетических конфигураций. Правильный выбор значений λ, γ и скорости изменения $\mu(t)$ сильно влияет на качество решения и, как правило, требует эмпирической оценки.

Общие свойства и различия. Методы SQA и SB итеративно обновляют параметры, направляя систему к оптимальному решению:

- **SQA** использует *стохастическое выборочное* моделирование для имитации квантовых переходов с настраиваемыми флуктуациями;
- **SB** использует детерминированную динамику, в которой бинарные состояния формируются за счет бифуркационных процессов.

Мы рассматриваем оба метода в качестве оракула «черного ящика» в методе обучения с подкреплением:

$$\mathbb{P}\Big(H_{\text{QUBO}}\left(\mathbf{x}^*\right) \leqslant \min_{\mathbf{x}} H_{\text{QUBO}}(\mathbf{x}) + \delta\Big) \geqslant 1 - \eta, \tag{12}$$

где $\mathbf{x}^*$ — решение, предлагаемое квантовым алгоритмом, δ — допустимая погрешность, а η — уровень доверия.

Сначала мы использовали квантовый отжиг для первичной выборки траекторий в алгоритме обучения с подкреплением на основе метода Монте-Карло, как показано на рис. 1, постепенно переходя к более сложным архитектурам. Для разработки и запуска компонентов, использующих квантовые вычисления, мы используем облачную платформу *Qonquester*, разработанную компанией *QDeep*.

Ограничения эффективности выборки в обучении с подкреплением на основе метода Монте-Карло

Основным ограничением большинства современных методов стохастической оптимизации является предположение о том, что все элементы выборки независимы и одинаково распределены. В обучении с подкреплением это предположение нарушается, поскольку элементы выборки

могут коррелировать в рамках одного марковского процесса принятия решений (*Markov Decision Process*, MDP). Кроме того, эффективная реализация большинства алгоритмов предполагает знание времени смешения и асимптотического поведения MDP. Для MDP с многомерным пространством состояний или разреженной структурой вознаграждений определение времени смещения часто невозможно [Wolfer, 2020], а при его наличии ограничивается практическая применимость таких методов.

Методы Монте-Карло (МС) оценивают функцию ценности на основе возвращаемых значений при выборе конкретного состояния. Однако эти оценки сильно зависят от выбранной политики, генерирующей траектории, и ничего не гарантирует, что выбранная политика оптимальна. Поэтому оценочная функция может не сойтись к оптимальному значению [Liu, 2021; Winnicki, Srikant, 2023].

Распространенным подходом для преодоления этих ограничений является стратегия «исследование – применение» (*exploration – exploitation*) [Sutton, Barto, 1998], при которой агент в разные моменты времени выполняет случайные действия для исследования новых состояний (exploration), а в остальных случаях принимает решения на основе текущей оценочной функции (exploitation). Данный подход позволяет уменьшить смещение, вызванное фиксированной политикой, однако он может сходиться медленно и вносить нестабильность в процесс обучения. Для состояния s значение оценочной функции методом Монте-Карло на основе N эпизодов определяется следующим образом:

$$\widehat{V}(s) = \frac{1}{N_s} \sum_{k=1}^{N} \mathbb{I}_k(s) G_k(s), \tag{13}$$

где $G_k(s)$ — суммарное вознаграждение, полученное в k-м эпизоде при посещении состояния s, $N_s = \sum\limits_{k=1}^{N} \mathbb{I}_k(s)$ — количество посещений состояния s, а $\mathbb{I}_k(s)$ — индикатор того, что состояние s посещалось в k-м эпизоде.

Такой подход требует больших значений N по следующим причинам.

1. *Распространение дисперсии.* Оценка демонстрирует накопление дисперсии:

$$\mathrm{Var}[G_k(s)] \geqslant \frac{\sigma^2}{1-\gamma^2}, \tag{14}$$

где σ^2 — дисперсия вознаграждения, а γ — дисконтный множитель. Для уменьшения ошибки до значения менее ϵ требуется $N_s > O\left(\epsilon^{-2}\right)$ эпизодов.

2. *Субоптимальное разбиение траектории.* Оценка разлагается следующим образом:

$$\widehat{V}(s) = \alpha \mathbb{E}[G \mid \mathcal{T}_{\mathrm{opt}}] + (1-\alpha)\mathbb{E}[\mathbb{E}[G \mid \mathcal{T}_{\mathrm{sub}}]], \tag{15}$$

где $\alpha = \frac{|\mathcal{T}_{\mathrm{opt}}|}{N_s}$. Для сходимости требуется $\alpha > 0$, что приводит к экспоненциальному росту числа эпизодов N в задачах с редким вознаграждением.

3. *Двойственность исследования и вычислений.* Сходимость к ϵ-оптимальной окрестности требует

$$N_s \geqslant \frac{\log\left(\frac{|\mathcal{S}|}{\delta}\right)}{2\epsilon^2(1-\gamma)^2} \tag{16}$$

согласно анализу вероятностно-приближенно корректного обучения (*Probably Approximately Correct*, PAC), что делает задачу непрактичной при больших пространствах состояний $\mathcal{S}$.

Следуя логике выбора наиболее репрезентативного подмножества эпизодов, естественно стремиться минимизировать среднеквадратичное отклонение оценки ценности, построенной по подвыборке, от истинной ценности. В теории выбор оптимальной подвыборки $\mathcal{E}^* \subset \mathcal{E}$ из m эпизодов ($m \ll N$) может помочь обойти указанные ограничения, решив задачу оптимизации:

$$\min_{\mathcal{E}^* \subset \mathcal{E}} \left\| \widehat{V}_{\mathcal{E}^*}(s) - V^{\pi}(s) \right\|_2 \quad \text{s. t.} \quad |\mathcal{E}^*| = m. \tag{17}$$

Эта задача изоморфна задаче выбора признаков (*feature selection*), где эпизоды выступают в роли базисных функций. Однако такая оптимизация обладает следующими особенностями.

1. *Комбинаторный взрыв.* Пространство решений имеет $\binom{N}{m}$ конфигураций, что делает задачу неразрешимой при $N > 50$.

2. *Нелинейная взаимосвязь.* Точность оценки зависит от взаимодействия траекторий:

$$\Delta\widehat{V} = \sum_{i \in \mathcal{E}^*} w_i G_i + \sum_{i \neq j \in \mathcal{E}^*} w_i G_i G_j + O\left(G^3\right). \tag{18}$$

3. *NP-трудность.* Задача сводится к задаче максимального покрытия:

$$\max_{\mathbf{x}} \sum_{s \in \mathcal{S}} \min\left(1, \sum_i x_i \mathcal{I}_i(s)\right) \quad \text{s. t.} \quad \sum_i x_i = m, \quad x_i \in \{0, 1\}, \tag{19}$$

которая является NP-трудной через редукцию из задачи покрытия множеств.

Как было отмечено в § 2, квантовые вычисления и квантовые алгоритмы предоставляют перспективный инструментарий для решения подобных комбинаторных задач. Задача выбора переформулируется в виде задачи QUBO:

$$\mathbf{x}^* = \arg\min_{\mathbf{x} \in \{0, 1\}^N} \left(-\mathbf{c}^{\top}\mathbf{x} + \mathbf{x}^{\top}\mathbf{Q}\mathbf{x}\right), \tag{20}$$

где $\mathbf{c}$ содержит вознаграждения эпизодов, а матрица $\mathbf{Q}$ штрафует за схожесть траекторий. Метод квантового отжига исследует это энергетическое пространство с ожидаемой сложностью $O\left(\exp\left(\sqrt{N}\right)\right)$, тогда как классические алгоритмы имеют сложность $O(\exp(N))$ для разреженных QUBO-задач.

Такой подход преобразует метод Монте-Карло в обучении с подкреплением от статистического усреднения к комбинаторной оптимизации. Далее мы формализуем QUBO-представление для квантовой обработки.

Обучение политики методом Монте-Карло посредством QUBO-выборки эпизодов

Мы исследуем стратегию обучения в стохастической среде GridWorld, в которой агент собирает все выбранные эпизоды и обновляет таблицу значений $Q(s, a)$ с помощью метода Монте-Карло. Рассматриваются два подхода:

(i) классический метод Монте-Карло (MC), который использует все элементы выборки для каждого обновления стратегии (см. алгоритм 1);

(ii) метод Монте-Карло с интеграцией QUBO-оптимизации (MC + QUBO), включающий этап комбинаторной фильтрации между выборкой эпизодов и обновлением стратегии (см. алгоритм 2).

В методе MC + QUBO для n эпизодов $\{\tau_i\}_{i=1}^n$ $\left($каждый из которых содержит состояние S_i и накопленное вознаграждение $R_i = \sum\limits_t r_t(\tau_i)\right)$ решается задача квадратичной неограниченной бинарной оптимизации для выбора компактного подмножества эпизодов, обеспечивающего высокое суммарное вознаграждение и разнообразие состояний, перед обновлением стратегии методом Монте-Карло.

Результат использования стандартного метода Монте-Карло представлен в алгоритме 1. Алгоритм MC + QUBO (алгоритм 2) отличается дополнительным шагом: после выбора набора эпизодов он приводит их к каноническому виду, вычисляет вознаграждения эпизодов R_i и их попарную схожесть $w_{ij} \in [0, 1]$ $\left($например, коэффициент Жаккара $w_{ij} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}\right)$, затем формулирует и передает в квантовый солвер следующую задачу QUBO:

$$\min_{x \in \{0,1\}^n} \sum_{i=1}^{n} (-\alpha R_i) x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \gamma w_{ij} x_i x_j, \tag{21}$$

где $x_i = 1$ соответствует эпизоду τ_i, $\alpha > 0$ — регулирующий множитель вознаграждения, а $\gamma \geqslant 0$ — коэффициент штрафа за схожесть эпизодов.

В терминах QUBO-оптимизатора задача может быть представлена как

$$x^\top Q_{\text{qubo}} x, \tag{22}$$

где $(Q_{\text{qubo}})_{ii} = -\alpha R_i$, а $(Q_{\text{qubo}})_{ij} = \frac{1}{2}\gamma w_{ij}$ для $i \neq j$ (в зависимости от выбранных соглашений конкретного оптимизатора).

Уравнение (21) реализует компромисс между вознаграждением и разнообразием: линейный член $-\alpha R_i$ поощряет эпизоды с большим вознаграждением (минимизация уменьшает целевое значение при больших R_i), тогда как квадратичный член $\gamma w_{ij} x_i x_j$ штрафует за выбор схожих эпизодов, стимулируя разнообразие.

Множитель α используется для увеличения энергетического разрыва между качественными и некачественными эпизодами, чтобы штраф за схожесть не доминировал и оптимизатор выбирал эпизоды с меньшим вознаграждением, но более разнообразными состояниями. Эффективная эвристика для выбора α на практике имеет вид

$$\alpha \cdot \mathbb{E}[R] \gtrsim \gamma \cdot (k-1) \cdot \mathbb{E}[w], \tag{23}$$

где k — ожидаемый размер выбранного подмножества. Такое условие гарантирует, что линейное слагаемое вознаграждения репрезентативного положительного эпизода обычно превосходит дополнительный квадратичный штраф за схожесть при добавлении этого эпизода в текущее множество.

Задача QUBO передается в квантовый оптимизатор, который возвращает набор кандидатов $\left\{x^{(s)}\right\}$ вместе с их энергиями (целевыми значениями). Стандартный подход заключается в следующем: получить множество вариантов, отсортировать их по возрастанию энергии и выбрать наилучший эпизод:

$$x^* = \arg\min_s \text{energy}\left(x^{(s)}\right) \tag{24}$$

или выбрать набор топ-эпизодов с наименьшими энергиями для увеличения устойчивости решения. Индексы

$$\mathcal{I} = \left\{i \colon x_i^* = 1\right\} \tag{25}$$

определяют подмножество эпизодов, используемых для обновления политики методом Монте-Карло. Такой рабочий процесс выборки и отбора показан в алгоритме 2 (см. строки, в которых происходит отбор лучших элементов выборки).

Алгоритмы 1 и 2 представлены в виде псевдокода. Алгоритм 1 является опорным и использует все эпизоды, тогда как алгоритм 2 включает создание QUBO, обращение к квантовому оптимизатору, шаг отбора (выбор наилучших элементов выборки) и последовательное обновление методом Монте-Карло на отобранном подмножестве эпизодов.

Algorithm 1. Метод Монте-Карло (используются все эпизоды)

Вход: число выборок N, размер выборки m, среда, вероятность неуспеха
Инициализация политики π; $Q(s, a) \leftarrow 0$; пустая история возвратов
for выборка = 1 до N **do**
  Сгенерировать m эпизодов $\{\tau_1, \ldots, \tau_m\}$, следуя стратегии π (с учетом неуспеха)
  for каждый эпизод τ в батче **do**
    for каждое первое посещение (s, a) в τ **do**
      Вычислить возврат G (сумма будущих вознаграждений)
      Добавить G в returns(s, a)
      Обновить $Q(s, a) \leftarrow$ mean(returns(s, a))
    end for
  end for
  Обновить стратегию: $\pi(s) \leftarrow \arg\max_a Q(s, a)$ для всех состояний s
end for

Algorithm 2. Метод Монте-Карло + QUBO (отбор лучших эпизодов)

Вход: число выборок N, размер выборки m, вес вознаграждения α, вес штрафа за схожесть γ, QUBO-решатель, ожидаемый размер подвыборки k
Инициализировать политику π; $Q(s, a) \leftarrow 0$; пустая история возвратов
for выборка = 1 до N **do**
  Сгенерировать m эпизодов $\{\tau_1, \ldots, \tau_m\}$, следуя стратегии π
  Опционально привести эпизоды к каноническому виду (первое посещение)
  Вычислить $R_i \leftarrow$ суммарное вознаграждение эпизода τ_i и множество посещенных состояний S_i, $i = 1, \ldots, m$
  Вычислить попарные сходства $w_{ij} \in [0, 1]$ (например, коэффициент Жаккара) для всех $i < j$
  Сформировать коэффициенты QUBO: линейные $h_i = -\alpha R_i$, квадратичные $J_{ij} = \gamma w_{ij}$
  Передать QUBO (h, J) в решатель; запросить множество образцов
  Получить набор образцов $\left\{x^{(s)}\right\}$ с энергиями; выбрать лучший $x^* = \arg\min_s \text{energy}\left(x^{(s)}\right)$
  Опционально обработать топ-образцы (голосование/ансамбль) для устойчивого x^*
  Пусть $\mathcal{I} = \left\{i\colon x_i^* = 1\right\}$; выбранные эпизоды $\{\tau_i\}_{i \in \mathcal{I}}$
  for каждый выбранный эпизод τ_i **do**
    for каждое первое посещение (s, a) в τ_i **do**
      Вычислить возврат G; добавить в returns(s, a); обновить $Q(s, a)$
    end for
  end for
  Обновить стратегию: $\pi(s) \leftarrow \arg\max_a Q(s, a)$
end for

Эксперимент и результаты

Мы оценили предложенный алгоритм отбора эпизодов MC + QUBO в средах GridWorld с ограниченным горизонтом и размерами $\{3 \times 3, 5 \times 5, 8 \times 8, 10 \times 10, 15 \times 15, 20 \times 20\}$, сравнивая его с базовым методом Монте-Карло. В каждом эксперименте агент обучался на нескольких наборах эпизодов, а качество оценки проводилось по следующим метрикам: скорость сходимости, качество итоговой стратегии, стабильность обучения.

Во всех протестированных конфигурациях алгоритм MC + QUBO демонстрировал сходимость при меньшем размере выборки, при этом эффект усиливался с увеличением размера среды ($\geqslant 10 \times 10$), где редкие вознаграждения и экспоненциальный рост пространства состояний обычно затрудняют оценку политики. Интеграция метода отбора эпизодов на основе задачи QUBO позволила алгоритму фокусироваться на наиболее информативных и разнообразных траекториях, избегая избыточных эпизодов, которые не вносят новой информации для функции ценности.

Такой отбор формулируется в виде задачи QUBO, решаемой с помощью квантового оптимизатора, что обеспечивает получение решений в пределах практических вычислительных затрат.

Ключевым элементом нашей реализации является то, что мы не оптимизируем величину вознаграждения напрямую. Несмотря на то что формулировка задачи QUBO включает член, связанный с вознаграждением, в основных экспериментах его влияние было сведено к нулю, чтобы отдать приоритет охвату пространства состояний.

Такой подход снижает риск переобучения на малой выборке эпизодов с высокими вознаграждениями, которое могло бы сместить обновление стратегии и сократить исследование среды. Предварительные эксперименты показали, что ориентация на вознаграждение приводит к быстрой, но нестабильной сходимости, тогда как наша формулировка, независимая от вознаграждения, обеспечивает *более сбалансированное обучение и лучшую обобщающую способность*.

Итоговые стратегии, представленные на рис. 2, во всех средах достигли лучших средних значений вознаграждений по сравнению со стратегиями, полученными стандартным методом Монте-Карло. Наибольший прирост качества наблюдался в более крупных сетках, где MC + QUBO сохранял разнообразие исследовательского поведения при одновременном ускорении сходимости.

Более того, масштабируемость нашего подхода на основе QUBO демонстрирует, что квантовая оптимизация может быть естественно интегрирована в качестве подпроцесса принятия решений в обучении с подкреплением.

Обсуждение и замечания

Полученные результаты демонстрируют стабильное превосходство метода MC + QUBO над стандартным методом Монте-Карло, особенно на больших сетках ($\geqslant 10 \times 10$). Отбор эпизодов с помошью квантового отжига позволяет преодолеть ключевые ограничения метода Монте-Карло:

- избыточность выборки устраняется с помощью квадратичного штрафа за схожесть эпизодов;
- размывание редких вознаграждений компенсируется за счет приоритета на охват пространства состояний;
- накопление дисперсии уменьшается за счет фильтрации скоррелированных эпизодов.

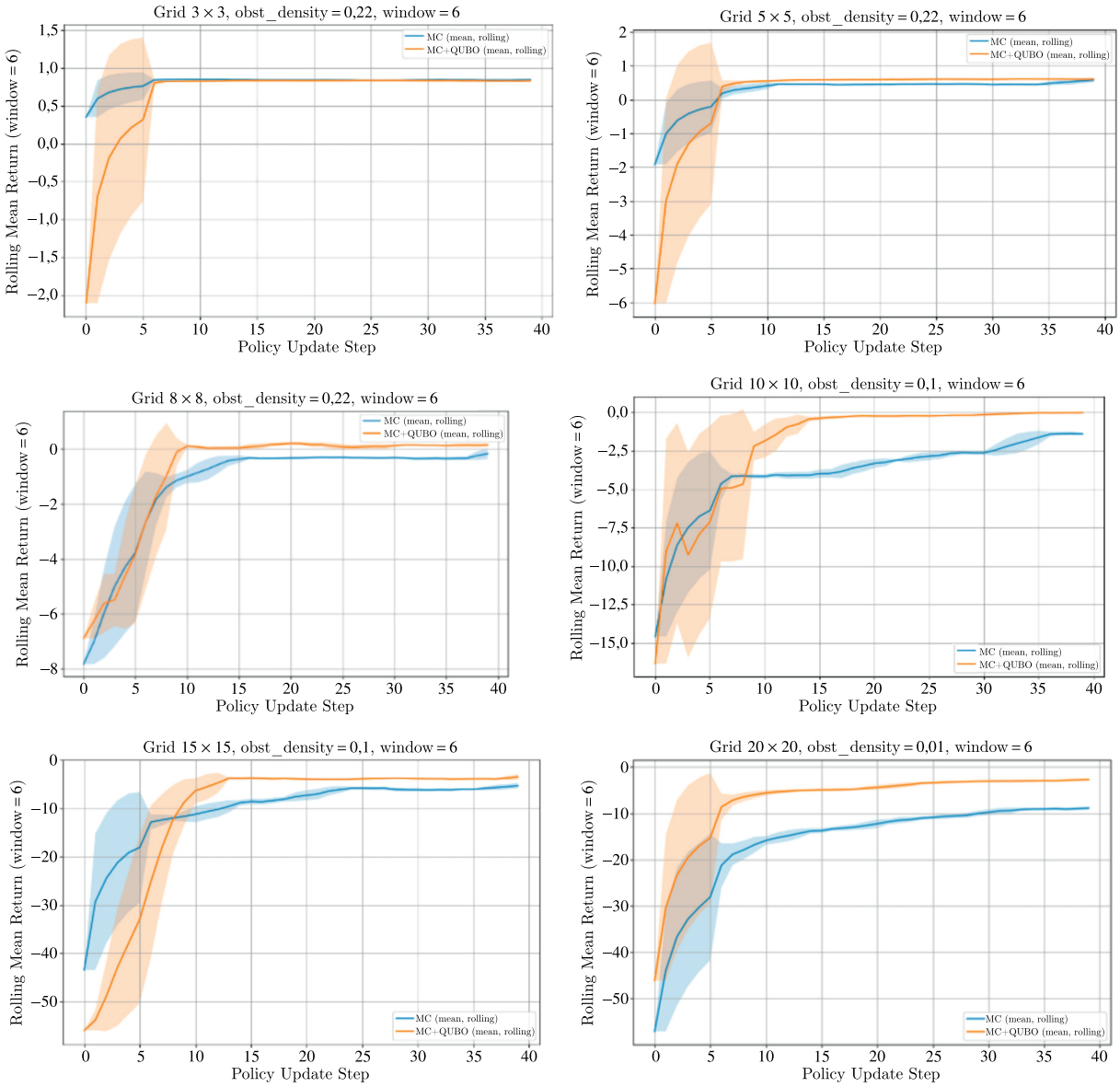

Рис. 2. Скользящее среднее вознаграждений (окно сглаживания размером 6) по обновлениям стратегии методом Монте-Карло и комбинированным методом Монте-Карло с QUBO в среде GridWorld. Верхний ряд показывает результаты для малых сеток (3×3, 5×5, 8×8) с заградительной плотностью 0,22. Нижний ряд содержит результаты для больших сеток (10×10, 15×15) с заградительной плотностью 0,1 и для сетки 20×20 с заградительной плотностью 0,01. Голубые линии соответствуют методу Монте-Карло (MC), оранжевые линии — комбинированному методу MC + QUBO, а заштрихованные области отражают вариацию вознаграждений

Реализация. Мы использовали облачную платформу Qonquester Cloud для выполнения вычислений методом SB и его тепловых вариаций [Kanao, Goto, 2022], а также симулятора квантового отжига (SQA). Квантовые оптимизаторы продемонстрировали превосходство над классическими подходами (такими как поиск с запретами и имитационный отжиг), однако эти результаты опущены для краткости.

Вычисления. Задержка при решении QUBO-задачи (~ 0,5–2,0 с/пакет) в основном обусловлена сетевой задержкой при передаче матриц. Реальное время работы оптимизатора пренебрежимо мало (~ 10–100 мс) при $n \leqslant 200$. Причем это время было измерено для облачного API

Qonquester, а при локальном запуске с оптимизированными реализациями SQA/SB возможна обработка и существенно больших n.

Заключение

Наши результаты подтверждают, что QUBO-отбор эпизодов представляет собой эффективный и практичный инструмент для улучшения обучения с подкреплением на основе метода Монте-Карло, особенно в средах с редкими вознаграждениями, высокой размерностью состояний или скоррелированными траекториями.

Помимо задач в дискретной постановке, предложенный метод может быть расширен на задачи непрерывного управления, иерархическое обучение с подкреплением и мультиагентные системы. Перспективными направлениями дальнейших исследований являются динамическая настройка весов отбора, гибридные критерии выбора эпизодов и развертывание алгоритма на реальных квантовых компьютерах. В задачах обучения с подкреплением предложенный метод может быть использован для отбора релевантных переходов из replay buffer, в которых количество элементов может быть ограничено размером буфера или мини батча.

Связывая обучение с подкреплением с комбинаторной оптимизацией, мы открываем возможности для нового класса алгоритмов, в которых процесс обучения управляется квантовыми или квантово-вдохновленными методами. При этом необходимо всегда искать компромисс между качеством отбора и вычислительными затратами при росте размерности.

Список литературы (References)

Bashar M., Shukla N. Designing Ising machines with higher order spin interactions and their application in solving combinatorial optimization // Scientific Reports. — 2023. — Vol. 13. — P. 9558.

Bertsekas D. P., Tsitsiklis J. N. Neuro-dynamic programming. — Belmont, MA: Athena Scientific, 1996. — DOI: 10.1007/978-0-387-74759-0_440

Cipra B. A. An introduction to the Ising model // Am. Math. Monthly. — 1987. — Vol. 94, No. 10. — P. 937–959. — https://doi.org/10.2307/2322600

Crosson E., Harrow A. Simulated quantum annealing can be exponentially faster than classical simulated annealing // 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS). — 2016. — P. 714–723. — DOI: 10.1109/FOCS.2016.81

Das A., Kempe D. Submodular meets spectral: greedy algorithms for subset selection, sparse approximation and dictionary selection // Proceedings of the 28th International Conference on Machine Learning (ICML 2011). — Bellevue, Washington, USA: Omnipress, 2011. — P. 1057–1064. — DOI: 10.48550/arXiv.1102.3975

Feige U. A threshold of $\ln n$ for approximating set cover // Journal of the ACM. — 1998. — Vol. 45, No. 4. — P. 634–652. — DOI: 10.1145/285055.285059

Garey M. R., Johnson D. S. Computers and intractability: a guide to the theory of NP-completeness. — San Francisco, CA: W. H. Freeman and Company, 1979.

Glover F., Kochenberger G. A. Quantum bridge analytics I: a tutorial on formulating and using QUBO models // Annals of Operations Research. — 2022. — Vol. 314. — P. 141–183. — DOI: 10.1007/s10479-022-04634-2

Goemans M. X., Williamson D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming // J. ACM. — 1995. — Vol. 42, No. 6. — P. 1115–1145. — DOI: 10.1145/227683.227684

Goto H., Tatsumura K., Dixon A. R. Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems // Sci. Adv. — 2019. — Vol. 5, No. 4. — P. eaav2372. — DOI: 10.1126/sciadv.aav2372

Greensmith E., Bartlett P. L., Baxter J. Variance reduction techniques for gradient estimates in reinforcement learning // Journal of Machine Learning Research. — 2004. — Vol. 5. — P. 1471–1530.

Guyon I., Elisseeff A. An introduction to variable and feature selection // Journal of Machine Learning Research. — 2003. — Vol. 3. — P. 1157–1182.

Hauke P., Bonnes L., Buyskikh A. S., Cosme J., de Lénica P., Lahaye T., Lewenstein M., Lyu Y., Nigmatullin R., Salathé Y., Tagliacozzo L., Viglietta E. Perspectives of quantum annealing: methods and implementations // Reports on Progress in Physics. — 2020. — Vol. 83. — P. 054401. — DOI: 10.1088/1361-6633/ab85b8

Jiang N., Li L. Doubly robust off-policy evaluation for reinforcement learning // Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). — 2016. — P. 652–661.

Liu J. On the convergence of reinforcement learning with Monte Carlo exploring starts // Automatica. — 2021. — Vol. 129. — P. 109693. — DOI: 10.1016/j.automatica.2021.109693

Kadowaki T., Nishimori H. Quantum annealing in the transverse Ising model // Physical Review E. — 1998. — Vol. 58. — P. 5355–5363. — DOI: 10.1103/PhysRevE.58.5355

Kanao T., Goto H. Simulated bifurcation assisted by thermal fluctuation // Commun Phys. — 2022. — Vol. 5. — P. 153. — DOI: 10.1038/s42005-022-00929-9

Kearns M., Singh S. Finite-sample convergence rates for Q-learning and Monte Carlo methods // Advances in Neural Information Processing Systems 11 (NeurIPS 1998). — 2000. — P. 1005–1011.

Kochenberger G. A., Hao J.-K., Glover F., Lewis-Pye A., Wang H., Sima C., Gutjahr W. J. The unconstrained binary quadratic programming problem: a survey // Journal of Combinatorial Optimization. — 2014. — Vol. 28. — P. 54–81. — DOI: 10.1007/s10878-014-9734-0

Krause A., Golovin D. Submodular function maximization // Tractability: practical approaches to hard problems. — Cambridge University Press, 2014. — P. 71–104. — DOI: 10.1017/CBO9781139177801.004

Lucas A. Ising formulations of many NP problems // Frontiers in Physics. — 2014. — Vol. 2. — Article 5. — DOI: 10.3389/fphy.2014.00005

McGeoch C. C. Adiabatic quantum computation and quantum annealing: theory and practice // Synthesis Lectures on Quantum Computing. — 2014. — Vol. 8, No. 2. — DOI: 10.2200/S00585ED1V01Y201407QMC008

Morita S., Nishimori H. Mathematical foundation of quantum annealing // Journal of Mathematical Physics. — 2008. — Vol. 49. — P. 125210. — DOI: 10.1063/1.2995837

Nemhauser G. L., Wolsey L. A. Integer and combinatorial optimization. — New York, NY: Wiley, 1988. — DOI: 10.1002/9781118627372

Puterman M. L. Markov decision processes: discrete stochastic dynamic programming. — New York, NY: Wiley, 1994. — DOI: 10.1002/9780470316887

Salloum H., Zhanalin S., Al Badr A., Kholodov Y. Mini-scale traffic flow optimization: an iterative QUBOs approach converting from hybrid solver to pure quantum processing unit // Scientific Reports. — 2025. — Vol. 15. — Article 22904. — DOI: 10.1038/s41598-025-04568-2

Singh S., Sutton R. S. Reinforcement learning with replacing eligibility traces // Machine Learning Journal. — 1996. — Vol. 22, No. 1–3. — P. 123–158. — DOI: 10.1007/BF00114726

Sutton R. S., Barto A. G. Reinforcement learning: an introduction. — Cambridge, MA: MIT Press, 1998.

Thomas P. S., Brunskill E. Data-efficient off-policy policy evaluation for reinforcement learning // Proceedings of the 33rd International Conference on Machine Learning (ICML 2016). — 2016. — P. 1329–1338. — DOI: 10.48550/arXiv.1604.00923

Winnicki A., Srikant R. On the convergence of policy iteration-based reinforcement learning with Monte Carlo policy evaluation // Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS 2023). — Proc. Machine Learning Research, 2023. — Vol. 206. — P. 9852–9878.

Wolfer G. Mixing time estimation in ergodic Markov chains from a single trajectory with contraction methods // Proceedings of the 31st International Conference on Algorithmic Learning Theory. — PMLR, 2020. — Vol. 117. — P. 890–905. — https://proceedings.mlr.press/v117/wolfer20a.html

Zlokapa A., Carleo G. Quantum-inspired optimization with applications and limits // PRX Quantum. — 2021. — Vol. 2. — P. 040101. — DOI: 10.1103/PRXQuantum.2.040101