УДК 004.852+336.761

Глубокое обучение с подкреплением для оптимизации инвестиционного портфеля: применение в управлении активами энергетического сектора

Н.М. Кошелев, А.В. Тарлыков, А.П. Преображенский

Воронежский институт высоких технологий, Воронеж, Россия

В данной работе проводится рассмотрение применения глубокого обучения с подкреплением в задачах динамической оптимизации инвестиционного портфеля применительно к активам энергетического сектора. Задача управления портфелем формализована как марковский процесс принятия решений (MDP). Подробным образом анализируются алгоритмы PPO, DDPG и SAC с акцентом на механику, стоящую за математическими различиями. В ходе рассмотрения показано, что не существует универсально лучшего алгоритма: A2C и PPO систематически превосходят SAC на трендовых рынках (накопленный прирост +12,5% против +4,5%), тогда как SAC лидирует в периоды кризисной волатильности (коэффициент Шарпа 1,18 против 0,61 у Buy & Hold; максимальная просадка −19,3% против −38,2%). Вместе с тем рассматриваются вопросы интерпретируемости посредством SHAP и LIME, нестационарность финансовой среды и практические барьеры между бэктестированием и реальной торговлей.

Ключевые слова: глубокое обучение с подкреплением, оптимизация портфеля, марковский процесс принятия решений, PPO, DDPG, SAC, режимная зависимость.

Deep Reinforcement Learning for Investment Portfolio Optimization: Application in Asset Management of the Energy Sector

N.M. Koshelev, A.V. Tarlykov, A.P. Preobrazhenskiy

Voronezh Institute of High Technologies, Voronezh, Russia

This paper examines the application of deep reinforcement learning in the tasks of dynamic optimization of the investment portfolio in relation to the assets of the energy sector. The portfolio management problem is formalized as a Markov Decision Process (MDP). The algorithms PPO, DDPG, and SAC are analyzed with emphasis on the mechanics behind their mathematical differences. The central finding is that no algorithm is universally superior: A2C and PPO consistently outperform SAC in trending markets (cumulative return +12.5% vs. +4.5%), while SAC leads during high-volatility crises (Sharpe ratio 1.18 vs. 0.61 for Buy & Hold; max drawdown −19.3% vs. −38.2%). Interpretability via SHAP and LIME, financial environment non-stationarity, and practical barriers between backtesting and live trading are also discussed.

Keywords: deep reinforcement learning, portfolio optimization, Markov decision process, PPO, DDPG, SAC, regime dependence.

Введение

В феврале 2020 года корреляции между активами в большинстве диверсифицированных портфелей стремительно выросли к единице – за несколько недель до официального объявления пандемии. Диверсификация, на которой основана классическая портфельная теория, утрачивала защитные свойства именно тогда, когда они были наиболее необходимы.

Глубокое обучение с подкреплением (DRL) предлагает принципиально иной подход к управлению портфелем [1]: не рассчитывать оптимальное распределение активов из статичных параметров, а обучить агента адаптивно управлять этим распределением через взаимодействие со средой. Агент не предполагает нормальности распределений и не требует стабильной матрицы ковариаций – он обучается на опыте, включая кризисные эпизоды, при их наличии в обучающей выборке. Обзор подходов к применению DRL в финансовых приложениях проведён в работе [2].

Необходимо сразу сформулировать центральный вывод данной работы: результаты DRL-алгоритмов зависят от рыночного режима значительно сильнее, чем обычно признаётся в исследовательских публикациях. Один и тот же алгоритм может демонстрировать высокую эффективность на трендовом рынке и неудовлетворительную – при резкой смене режима. Данный вывод следует принимать во внимание при рассмотрении последующих разделов работы [2].

Почему модель Марковица не работает в энергетике

Классическая модель Марковица решает задачу оптимального распределения активов:

$$\max_{w} \mu^T w - \frac{\lambda}{2} w^T \sum w,\ w^T 1 = 1,\ w \geq 0, \quad (1)$$

где w – веса активов, μ – ожидаемые доходности, Σ – матрица ковариаций, λ – параметр неприятия риска. Задача (1) является квадратичной и решается аналитически или стандартными численными методами. В этом состоит её основное преимущество – и одновременно главная уязвимость.

В энергетическом секторе предположения модели нарушаются систематически. Ожидаемые доходности μ нестабильны: доходность акции нефтяной компании существенно зависит от текущей цены нефти, которая сама по себе не поддаётся надёжному прогнозированию. Матрица ковариаций Σ является нестационарной. Наконец, модель (1) не учитывает транзакционные издержки – DRL-агент включает их непосредственно в функцию вознаграждения, что принципиально изменяет оптимальную частоту ребалансировки.

Марковский процесс принятия решений

Управление портфелем формализуется в виде кортежа (S, A, P, R, γ) [1]. Схема взаимодействия агента со средой в рамках MDP представлена на рисунке 1. Состояние S_t в момент t включает текущие веса портфеля, ценовые и объёмные характеристики активов за предшествующие k периодов и макроэкономические индикаторы:

$$S_t = \left(W_{t-1}, p_{t-k:t}, v_{t-k:t}, m_t\right). \quad (2)$$

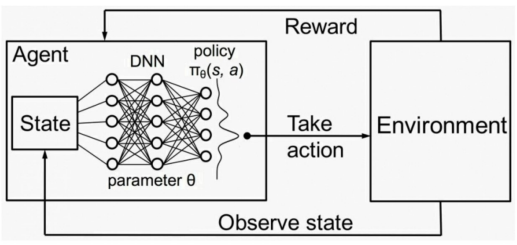

Рисунок 1. Схема взаимодействия агента со средой в рамках марковского процесса принятия решений

Действие a_t – целевой вектор весов после ребалансировки, принадлежащий единичному симплексу: $a_t \in \Delta N$. Функция вознаграждения с учётом транзакционных издержек с записывается следующим образом:

$$R(s_t, a_t) = log \frac{p_t^T a_t}{p_{t-1}^T a_{t-1}} - c \left\| a_t - a_{t-1} \right\|_1. \quad (3)$$

Первый член представляет собой логарифмическую доходность портфеля (логарифмическую ввиду аддитивности по времени). Второй – штраф за транзакционные издержки, пропорциональный L1-норме изменения весов. Целью агента является политика $\pi(a|s)$, максимизирующая кумулятивное дисконтированное вознаграждение:

$$J(\pi) = E_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right]. \quad (4)$$

Коэффициент $\gamma \in (0, 1)$ в выражении (4) регулирует горизонт планирования: при $\gamma \to 1$ агент ориентируется на долгосрочный результат. В задачах портфельного управления γ обычно выбирается в диапазоне 0,99–0,999.

PPO, DDPG, SAC: что реально стоит за формулами

Все три алгоритма относятся к классу актор-критик. Актор $\pi_\theta(a|s)$ задаёт политику; критик $V_\varphi(s)$ или $Q_\varphi(s, a)$ оценивает её качество.

PPO [3] обновляет политику через «обрезанную» суррогатную функцию потерь:

$$L^{CLIP}(\theta) = E \left[\min \left(r_t(\theta) \widehat{A}_t, \text{clip} \left(r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \widehat{A}_t \right) \right], \quad (5)$$

где $r_t(\theta) = \pi_\theta(a_t|s_t) / \pi_{\theta old}(a_t|s_t)$ – отношение новой и старой политик, $\varepsilon \approx 0{,}1–0{,}2$. Оператор clip ограничивает шаг обновления: политика не может измениться кардинально за одну итерацию. PPO является одним из наиболее стабильных алгоритмов именно благодаря данному свойству.

DDPG работает с детерминированной политикой $\mu_\theta(s)$. Критик $Q_\varphi(s, a)$ обучается минимизировать ошибку Беллмана:

$$L(\phi) = E \left[\left(Q_\phi(s_t, a_t) - \left(r_t + \gamma Q_{\phi'}\left(s_{t+1}, \mu_{\theta'}(s_{t+1}) \right) \right) \right)^2 \right]. \quad (6)$$

Актор же обновляется по градиенту Q-функции по действиям:

$$\nabla_\theta J \approx E \left[\nabla_a Q_\phi(s, a) \cdot \nabla_\theta \mu_\theta(s) \right]. \quad (7)$$

Детерминированность составляет сильную сторону DDPG в условиях стабильной среды и слабую – при нестационарности.

SAC [4] дополняет целевую функцию максимизацией энтропии политики:

$$J_{SAC}(\pi) = E_\pi \left[\sum_t \gamma^t \left(R(s_t, a_t) + \alpha H(\pi(\cdot|s_t)) \right) \right], \quad (8)$$

где $H(\pi(\cdot|s)) = -E[log \pi(a|s)]$ – энтропия политики, α – коэффициент температуры. Энтропийный член вынуждает агента сохранять стохастичность: вероятностная масса не концентрируется в одном действии. Это повышает робастность к нестационарности: агент с распределённой политикой быстрее переориентируется при смене режима.

Результаты: режимная зависимость как главный вывод

Эмпирические результаты [5] демонстрируют принципиальную зависимость производительности алгоритмов от рыночного режима. На рисунке 2 представлено сравнение накопленных вознаграждений для пяти алгоритмов в период декабрь 2023 – февраль 2024 года, характеризующийся выраженным трендовым движением.

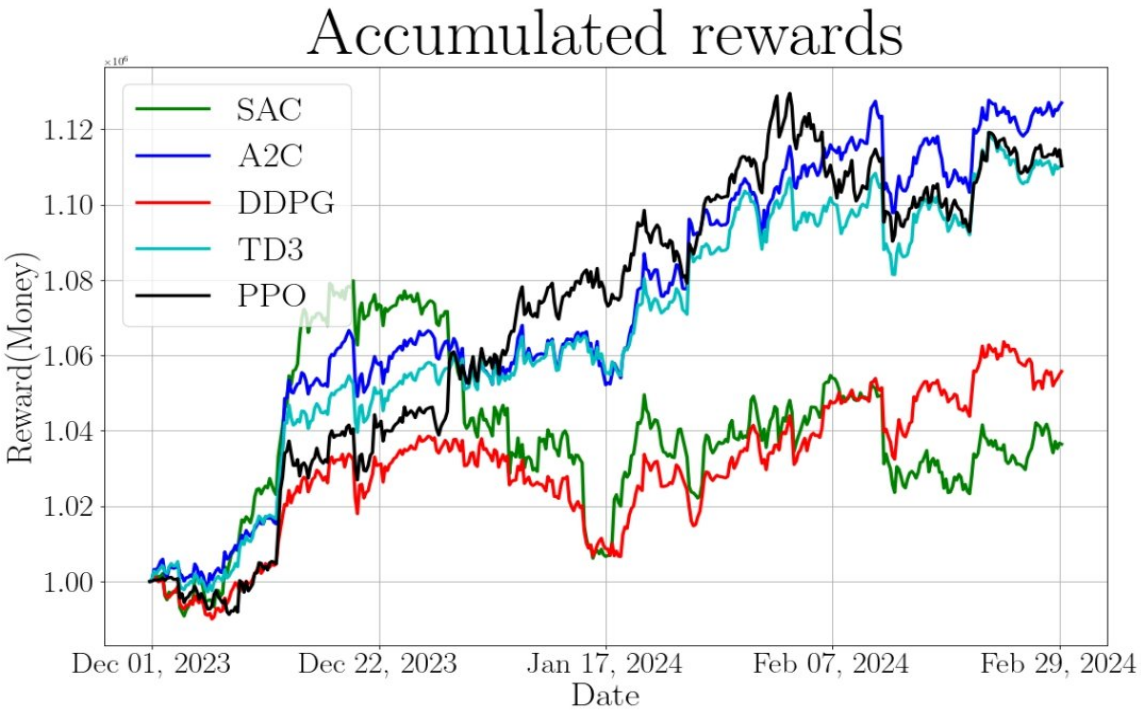

Рисунок 2. Накопленные вознаграждения алгоритмов DRL (SAC, A2C, DDPG, TD3, PPO) в период декабрь 2023 – февраль 2024 года

На трендовом рынке декабря 2023 – февраля 2024 года (рис. 2) A2C показал накопленный прирост около +12,5%, PPO – около +11,5%, тогда как SAC и DDPG – только +4–5%. Механика данного результата состоит в следующем: в условиях устойчивого восходящего тренда максимизация энтропии в SAC согласно выражению (8) порождает избыточную диверсификацию. Агент буквально наказывается за высокую уверенность в конкретном решении.

Картина кардинально меняется в период высокой волатильности 2022 года. SAC показал коэффициент Шарпа 1,18, максимальную просадку −19,3%, годовую доходность +22,4% и коэффициент Calmar 1,16. Для сравнения: PPO – Шарп 0,94, MaxDD −24,1%; DDPG – Шарп 0,71, MaxDD −31,4%; A2C – Шарп 0,58, MaxDD −33,8%; Buy & Hold – Шарп 0,61, MaxDD −38,2%. SAC лидировал именно потому, что свойство, снижавшее его эффективность на трендовом рынке, стало ценным в условиях турбулентности: стохастичность политики не позволяет агенту «застрять» в концентрированной позиции при резком развороте.

Данный результат формулирует принципиальный вывод для практики: применение любого DRL-алгоритма в реальном портфельном управлении без учёта текущего рыночного режима является методологической ошибкой. Следствием является потребность в адаптивных мета-системах, способных классифицировать режим и переключаться между алгоритмами [5].

Интерпретируемость: что SHAP говорит о поведении агента

Применение SHAP к PPO-агенту, управляющему портфелем энергетических акций [6], позволяет получить следующую картину. Наиболее влиятельным предиктором решений агента является изменение цены базового энергоносителя (Brent – для нефтяных компаний, Henry Hub – для газовых). Вторым по значимости выступает RSI выше 70 (сигнал перекупленности), коррелирующий с решением о сокращении позиции. Третьим – объём торгов: нетипично низкий объём в сочетании с высокой волатильностью ассоциируется у агента с повышенным риском разворота.

Агент демонстрирует асимметрию реакции: он существенно сильнее реагирует на сигналы к сокращению позиций, нежели к их наращиванию. Штраф через L1-норму в функции вознаграждения (3) кодирует неявное неприятие риска. Агент «выучил», что ошибки в сторону удержания убыточной позиции обходятся дороже пропущенного роста. Таким образом, дизайн функции вознаграждения является содержательным решением, определяющим, что агент считает «успехом» [6].

LIME-анализ эпизодов, в которых агент принимал решения, расходящиеся с рыночным консенсусом, выявил следующий паттерн: в ряде случаев агент опирался на нетипичные комбинации признаков – высокая волатильность плюс низкий объём плюс технический уровень поддержки, – предшествовавшие разворотам тренда. Открытым остаётся вопрос о том, являются ли данные паттерны устойчивыми закономерностями или артефактами обучающей выборки.

Ограничения

Нестационарность. Онлайн-дообучение частично решает данную проблему, однако порождает новую: риск «катастрофического забывания», при котором агент слишком быстро перезаписывает ранее усвоенные паттерны. Задача обеспечения баланса между адаптивностью и стабильностью остаётся открытой.

Дефицит данных. DRL-агентам требуются миллионы взаимодействий со средой. Для фондового рынка с 250 торговыми днями в году это соответствует тысячам лет рыночной истории. Синтетические данные частично решают проблему, однако вводят новый риск: агент, обученный на синтетических ценах, может переобучиться к особенностям генеративной модели, а не к реальным рыночным механизмам.

Разрыв бэктест/реальная торговля. Рыночное воздействие крупных ордеров, ограничения ликвидности, задержки исполнения не моделируются в стандартных бэктестах [5]. Игнорирование транзакционных издержек в формуле (3) может приводить к гиперактивным стратегиям, демонстрирующим высокую эффективность в симуляции, однако проигрывающим пассивным бенчмаркам в реальных условиях.

Заключение

Глубокое обучение с подкреплением обеспечивает реальную альтернативу статичной оптимизации Марковица (1) для управления портфелем в условиях нестационарных рынков. Формализация через MDP (2)–(4) обеспечивает строгую математическую основу; алгоритмы PPO (5), DDPG (6)–(7) и SAC (8) – практически применимые инструменты для непрерывных пространств действий.

Режимная зависимость алгоритмов является не техническим недостатком, устранимым улучшением архитектуры, а фундаментальным свойством, обусловленным природой самих алгоритмов. A2C и PPO систематически превосходят SAC на трендовых рынках; SAC лидирует в условиях кризисной волатильности (Шарп 1,18 против 0,61 у Buy & Hold, MaxDD −19,3% против −38,2%). Таким образом,

проведённое рассмотрение показало, что данный результат требует пересмотра устоявшейся практики сравнительного анализа DRL-алгоритмов в финансовой литературе.

Наиболее перспективными направлениями являются гибридизация DRL с трансформерным прогнозированием – агент, использующий вероятностные прогнозы TFT как часть вектора состояния (2), принципиально лучше оснащён для управления риском – и федеративное обучение для сценариев, в которых несколько фондов стремятся обучить общего агента без раскрытия состава портфелей [7]. Оба направления объединяют методы, рассмотренные в данном цикле работ.

СПИСОК ИСТОЧНИКОВ

1. Sutton R.S. Reinforcement Learning: An Introduction / R.S. Sutton, A.G. Barto. – 2nd ed. – Cambridge, MA: MIT Press, 2018. – 552 p.

2. A Review of Reinforcement Learning in Financial Applications / Y. Bai, Y. Gao, R. Wan [et al.] // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/2411.12746 (дата обращения: 25.02.2026).

3. Proximal Policy Optimization Algorithms / J. Schulman, F. Wolski, P. Dhariwal [et al.] // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/1707.06347 (дата обращения: 19.02.2026).

4. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor / T. Haarnoja, A. Zhou, P. Abbeel, S. Levine // Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018. – PMLR, 2018. – P. 1856–1865.

5. Deep Reinforcement Learning Strategies in Finance: Insights into Asset Holding, Trading Behavior, and Purchase Diversity / A. Mohammadshafie, A. Mirzaeinia, H. Jumakhan, A. Mirzaeinia // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/2407.09557 (дата обращения: 16.02.2026).

6. De-la-Rica-Escudero A. Explainable Post Hoc Portfolio Management Financial Policy of a Deep Reinforcement Learning Agent / A. de-la-Rica-Escudero, E.C. Garrido-Merchán, M. Coronado-Vaca // PLoS ONE. – 2025. – Vol. 20, No. 1. – URL: https://doi.org/10.1371/journal.pone.0315528 (дата обращения: 16.02.2026).

7. Ndikum Ph. Advancing Investment Frontiers: Industry-grade Deep Reinforcement Learning for Portfolio Optimization / Ph. Ndikum, S. Ndikum // arXiv [Электронный ресурс]. – URL: https://arxiv.org/abs/2403.07916 (дата обращения: 08.02.2026).

ИНФОРМАЦИЯ ОБ АВТОРАХ

Кошелев Никита Михайлович, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.

Тарлыков Александр Вячеславович, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.

Преображенский Андрей Петрович, доктор технических наук, профессор, Воронежский институт высоких технологий, Воронеж, Россия.