УДК 519.86:004.8:336.76

ПРИМЕНЕНИЕ МНОГОКРИТЕРИАЛЬНОЙ ФУНКЦИИ ВОЗНАГРАЖДЕНИЯ ДЛЯ ДИНАМИЧЕСКОГО УПРАВЛЕНИЯ РИСКОМ РЫНОЧНО-НЕЙТРАЛЬНЫХ ПОРТФЕЛЕЙ

Б. Е. Беляков

Московский институт электроники и математики им. А.Н. Тихонова НИУ ВШЭ;
Национальный исследовательский университет «Высшая школа экономики», г. Москва

✉ work.belyakov@gmail.com

Аннотация. Разработана и апробирована многокритериальная функция вознаграждения в рамках рыночно-нейтральной методологии управления инвестиционным портфелем на основе глубокого обучения с подкреплением, формализующая задачу оптимального распределения капитала как марковский процесс последовательных решений в нестационарной среде. Целевая функция вознаграждения одновременно оптимизирует избыточную доходность на единицу риска, нейтральность к рынку через введенный штраф за корреляцию с целевым рыночным индексом и трансакционные издержки. Таким образом, контроль системного риска и учет стоимости ребалансировки портфеля становятся частью процесса моделирования и обучения. Обучение и анализ модели выполняется при помощи кросс-валидации для снижения смещения и адаптации к смене режимов. Экспериментальная валидация на ведущих фондовых индексах США, Европы и Азии за десятилетний период показывает работоспособность предложенной методологии, в частности устойчивое улучшение метрик: более высокий коэффициент Шарпа, меньшую максимальную просадку портфеля и существенно сниженную корреляцию с рыночным индексом по сравнению с классическими алгоритмами оптимизации и подходами на основе обучения с подкреплением на основе одного критерия.

Ключевые слова: рыночно-нейтральная стратегия, оптимизация портфеля, обучение с подкреплением, рыночно-нейтральный портфель, глубокое обучение, управление портфелем.

ВВЕДЕНИЕ

Финансовые рынки характеризуются стохастичностью динамики доходностей при наличии устойчивых поведенческих и структурных аномалий, таких как импульс, избыточная волатильность и мультифакторные премии [1–4]. С учетом указанных характеристик рыночно-нейтральные стратегии сформировали ядро современной количественной торговли, где целью агента является извлечение избыточной доходности при минимальной экспозиции к систематическому риску, т. е. получение инвестиционного преимущества, не объяснимого пассивными рыночными факторами [2, 5, 6]. Исторически истоки данного подхода восходят к ранним хедж-фондам Альфреда Джонса (1940-е гг.), использовавшим конструкции из длинных и коротких позиций для хеджирования систематических рисков [7]. Теоретический фундамент заложен факторными моделями (Фамы – Френча, Кархарта) и классическими стратегиями относительной стоимости, включая торговлю парами или синтетическими контрактами [1, 2, 8].

При практической востребованности традиционные реализации подобных моделей опираются на статические факторные структуры с фиксированными экспозициями и подвержены ряду системных уязвимостей: деградации факторов при смене рыночных режимов [4, 9, 10], переобучению в высокоразмерных пространствах признаков [9, 11, 12] и отсутствию механизмов динамической адаптации к кластерам волатильности и макроэко-

номическим шокам [4, 10, 13]. Современные разработки в области тайминга факторных премий и адаптивного распределения капитала демонстрируют пользу динамических предикторов относительно статических арбитражных конструкций [11, 14], однако практическая польза таких моделей деградирует, когда целевая постановка задачи требует нулевой экспозиции к бенчмарку при реалистичных рыночных условиях и трансакционных издержках [3, 5, 14].

Глубокое обучение с подкреплением естественным образом формулирует управление портфелем как марковский процесс принятия решений, в котором агент обучается политике ребалансировки, оптимизирующей долгосрочную целевую функцию вознаграждения с учетом трансакционных издержек, ограничений на оборотный или заемный капитал, а также рисков просадки (снижения стоимости) стоимости портфеля [14–18]. Предлагаемая постановка задачи позволяет: обновлять представления по мере поступления данных [15, 17], эксплуатировать динамическую структуру доходностей и ковариаций [9, 11, 17], реализовывать замкнутый контур «наблюдение – действие – обратная связь», устраняя необходимость в жестко заданных наборах факторов [14, 15, 17].

Для практической реализации и стабильного обучения политик полезно опираться на современные алгоритмы оптимизации и техники регуляризации и предотвращения переобучения при высоких размерностях признаков [12, 16, 19].

Целью настоящей работы является разработка и эмпирическая оценка методологии для динамического рыночно-нейтрального распределения капитала на основе глубокого обучения с подкреплением (рис. 1). Управление портфелем формализуется как марковский процесс последовательных решений в нестационарной среде, где агент выбирает вектор весов портфеля с учетом ограничений на экспозицию (выделенную долю капитала в отдельный актив и в стратегию в целом) и оборот (изменение весов).

Ключевым элементом разработанной методологии является многокритериальная функция вознаграждения, специально сконструированная для одновременного учета трех взаимосвязанных критериев:

1) максимизации избыточной доходности на единицу волатильности портфеля;

2) минимизации сопряженности с целевым рыночным индексом через штраф за корреляцию;

3) ограничения торгового оборота через штраф за трансакционные издержки.

Таким образом, контроль риска, рыночной нейтральности и стоимости ребалансировки включается непосредственно в целевую функцию агента, а не реализуется в виде внешних постобработок или жестких ограничений, что ранее не рассматривалось в подобных задачах.

Архитектура методологии включает конвейер альфа-сигналов (предикторов избыточной доходности; показатели, связанные с будущей динамикой цен активов), гибридный модуль извлечения признаков на основе глубоких нейронных сетей и рекуррентный алгоритм актор – критик, отображающий скрытое состояние в вектор целевых весов [16, 17]. Оценка проводится при помощи кросс-валидации с ежедневной ребалансировкой портфеля, с учетом проскальзывания и комиссий, в том числе стоимости коротких позиций [5, 10, 13].

В совокупности это отличает работу от актуальных подходов, ориентированных преимущественно на максимизацию доходности агента без явного контроля нейтральности и торгового оборота [10, 13]. На данных ключевых региональных индексов (США, Европа, Азия) предлагаемый подход последовательно демонстрирует повышение коэффициента Шарпа, снижение максимальной просадки портфеля (снижения стоимости портфеля по сравнению с последним пиком) и поддержание низкой или нулевой корреляции с бенчмарками.

Научная новизна работы состоит в интеграции многокритериальной функции вознаграждения, одновременно включающей риск-скорректированную доходность, штраф за рыночную корреляцию и трансакционные издержки, в рекуррентную нейронную сеть агента для построения рыночно-нейтральных портфелей. Предложенная в работе постановка задачи и функция вознаграждения ранее систематически не рассматривались в контексте агентного моделирования и обучения с подкреплением для рыночно-нейтральных стратегий.

В последующих разделах представлены обзор литературы, формализация процесса принятия решений, подробное описание архитектуры и целевой функции, постановку эксперимента и метрики, результаты численных экспериментов и направления дальнейших исследований.

1. ОБЗОР ЛИТЕРАТУРЫ

1.1. Рыночно-нейтральные портфели

Истоки рыночно-нейтральных стратегий восходят к первым хедж-фондам А. Джонса, где прибыль формировалась из отбора акций при одновременном хеджировании рыночного риска через

длинные и короткие позиции [5, 7]. Теоретическая база опирается на многокомпонентные модели ценообразования активов: трехфакторную модель Фамы – Френча и ее расширение Кархарта за счет фактора импульса [1, 2, 6]. Практические методы включают регрессионную нейтрализацию факторов и оптимизационные постановки с нулевой экспозицией на систематические факторы [9, 13]. Классические стратегии относительной стоимости – парный трейдинг, арбитраж конвертируемых облигаций и арбитраж слияний – демонстрировали экономически значимые премии; для парного трейдинга на акциях США за 1962–2002 гг. показаны статистически устойчивые избыточные доходности порядка 11 % годовых [8, 20].

Развитие машинного обучения расширило инструментарий статистического арбитража: ансамблевые модели и глубокие нейронные сети применяются для построения длинно-коротких портфелей и демонстрируют положительные результаты в задачах эмпирического ценообразования и прогнозирования доходностей [10, 11, 21]. Методы глубокого обучения и обучения с подкреплением используются для совместного построения портфелей, моделирования и оптимизации торговых правил [10, 14, 15]. В регрессионных моделях доходности проецируются на отраслевые и рыночные факторы с последующей нейтрализацией факторных нагрузок (например, через факторную регрессию Фамы–Френча) [1, 2, 9]. В оптимизационных постановках критерий Марковица служит базой для формулировок с жесткими ограничениями на экспозиции и риски, что применимо и к портфелям малой капитализации при корректной оценке ковариаций в высоких размерностях [6, 9]. Понятие нейтральности неоднородно: различные операционные критерии (регрессионная нейтрализация, прямые ограничения на экспозиции или целевые оптимизационные функции) приводят к различным практическим результатам применения стратегий [2, 9, 13]. Альтернативные оптимизационные формулировки, нацеленные на снижение корреляции с бенчмарком или на контроль торгового оборота, потенциально дают более строгий контроль нейтральности по сравнению с регрессиями, однако сопоставление эффектов зависит от конкретной постановки задачи и данных и требует глубокой эмпирической валидации [13, 22]. Наконец, несмотря на существенный прогресс методов обучения с подкреплением в задачах управления портфелем и учета рыночных ограничений [10, 23–25], их применение к рыночно-нейтральным портфелям и интеграция в функцию вознаграждения остаются недостаточно проработанными в литературе.

1.2. Обучение с подкреплением в торговых задачах

Обучение с подкреплением противопоставляется традиционным эконометрическим подходам благодаря его способности к онлайновому обучению и использованию глубоких нейросетевых архитектур, что делает его особенно востребованным в условиях шума, нестационарности и тяжелых хвостов распределений финансовых временны́х рядов [10, 11, 14, 15, 26].

Сравнительные обзоры указывают на сопоставимую эффективность ключевых семейств алгоритмов (методы градиента политики, алгоритм актор – критик) при решении задач управления портфелем, причем выбор конкретного подхода определяется формулировкой задачи и структурными ограничениями [15, 16, 26]. На практике ряд исследований демонстрирует применимость алгоритмов глубоких нейронных сетей и их рекуррентных модификаций к частично наблюдаемым финансовым задачам, а также выявляет преимущества методов градиента политики и алгоритма актор – критик, включая их детерминистические и стохастические разновидности, например проксимальную оптимизацию политики, для оптимизации целевых функций портфеля [10, 16, 18, 19]. Следует отметить, что критическим фактором остаются спецификация марковского процесса принятия решений и выбор признакового пространства. Включение нефинансовых индикаторов, например показателей рыночных настроений, наряду с ценовыми характеристиками, по данным ряда исследований, способствует улучшению прогнозных свойств моделей и приводит к повышению показателей доходности и снижению рисковых характеристик в эмпирических экспериментах. Вместе с тем величина этого эффекта существенно зависит от состава признаков, выбранной схемы регуляризации и методов оценки ковариационных структур в пространствах высокой размерности [10, 12, 21, 26]. В задачах маркетмейкинга (поддержания ликвидности торгового инструмента путем регулярного выставления заявок на покупку и продажу) и многопрофильного распределения капитала обучение с подкреплением эффективно масштабируется на большие, включая непрерывные, пространства действий и снижает зависимость от ручной параметризации за счет глубоких политик и детерминистических схем обучения [16–18, 19, 26]. Риск-скорректированные функции вознаграждения позволяют формализовать баланс доходности и риска [22, 27], а учет трансакционных издержек и проскальзывания является ключевым условием корректной оценки [5, 13, 20]. Указанные соображения положены в основу разрабо-

танной методологической схемы, которая формирует новый подход к обучению агента рыночно-нейтральным стратегиям.

2. МЕТОДОЛОГИЯ

2.1. Основные обозначения

Рассматривается дискретное время $t = 0, 1, ..., T$ и множество активов из N финансовых инструментов. Обозначим через

$$r_{t+1} = \left(r_{t+1}^1, ..., r_{t+1}^N \right)^{\mathrm{T}}$$

вектор доходностей активов за период $[t, t+1]$. Доходность выбранного рыночного индекса (бенчмарка) за тот же период обозначим через m_{t+1}.

Портфель в момент времени t описывается вектором весов

$$w_t = \left(w_t^1, ..., w_t^N \right)^{\mathrm{T}},$$

где w_t^i интерпретируется как позиционный вес, экспозиция i-го инструмента, выраженная в доле капитала, выделенной в актив. Так, $w_t^i > 0$ соответствует длинной позиции, $w_t^i < 0$ – короткой, $w_t^i = 0$ – отсутствию позиции. В отличие от классического полностью инвестированного портфеля с бюджетным ограничением $\sum_{i=1}^{N} w_t^i = 1$, в рыночно-нейтральной постановке задачи масштаб портфеля фиксируется ограничением на валовую экспозицию (см. формулу (2) ниже).

Требование рыночно-нейтральной нетто-экспозиции портфеля формализуется условием

$$\sum_{i=1}^{N} w_t^i = 0, \qquad (1)$$

а ограничение на валовую (брутто-) экспозицию портфеля – неравенством

$$\sum_{i=1}^{N} \left| w_t^i \right| \le \Gamma, \qquad (2)$$

где $\Gamma > 0$ – заданный порог максимальной совокупной позиции.

Ограничение (2) задает верхнюю границу валовой экспозиции портфеля, т. е. ограничение на заемный капитал. В стратегиях порог Γ отражает максимально допустимую брутто-экспозицию (сумму абсолютных значений длинных и коротких позиций) и служит ограничением на заемный капитал. В рамках данной работы порог Γ фиксирован на уровне 100 %, что соответствует консервативному ограничению на совокупную позицию и обеспечивает нулевую нетто-экспозицию.

Совместно с условием (1) это означает, что суммарная длинная экспозиция равна суммарной короткой (в абсолютном выражении) и каждая из них не превышает $\Gamma/2$.

Изменение весов при ребалансировке портфеля в момент t равно

$$\Delta w_t = w_t - w_{t-1}.$$

Трансакционные издержки TC_t задаются в виде линейно-квадратичной функции от изменения портфеля:

$$TC_t = c_{\mathrm{lin}} \sum_{i=1}^{N} \left| \Delta w_t^i \right| + c_{\mathrm{quad}} \sum_{i=1}^{N} \left(\Delta w_t^i \right)^2, \qquad (3)$$

где $c_{\mathrm{lin}}, c_{\mathrm{quad}} > 0$ – параметры комиссий и проскальзывания.

Доходность портфеля p (приращение капитала на единицу выделенного капитала) с учетом издержек за период $[t, t+1]$ определяется как

$$r_{p,t+1} = \sum_{i=1}^{N} w_t^i r_{t+1}^i - TC_t. \qquad (4)$$

Формула (4) записана для указанных позиционных весов.

Если q_t^i обозначает денежный размер позиции, а C_t – капитал стратегии, то $w_t^i = \dfrac{q_t^i}{C_t}$ и $\dfrac{\sum_{i=1}^{N} w_t^i r_{t+1}^i}{C_t} = \sum_{i=1}^{N} w_t^i r_{t+1}^i$.

Таким образом, $r_{p,t+1}$ является доходностью на капитал стратегии, тогда как нетто-номинал $\sum_{i=1}^{N} q_t^i$ в рыночно-нейтральном портфеле равен нулю.

Трансакционные издержки TC_t в выражении (3) также заданы в долях капитала (через Δw_t) и поэтому вычитаются непосредственно в формуле (4).

Текущая рыночная сопряженность портфеля оценивается через коэффициент корреляции доходности портфеля с доходностью бенчмарка на скользящем окне длиной L_{corr}:

$$\rho_t = \mathrm{Corr}\left(r_{p,t+1},\, m_{t+1}\right) \qquad (5)$$

по данным на интервале $[t - L_{\mathrm{corr}} + 1,\, t+1]$.

Вектор наблюдаемых признаков среды в момент времени t обозначим через

$$x_t \in R^d,$$

где d – число показателей; в него входят ценовые, макроэкономические и альтернативные переменные.

2.2. Постановка задачи в терминах обучения с подкреплением

Обучение с подкреплением – парадигма оптимального управления, в которой агент взаимодействует со средой во времени, выбирая действия и получая оценки качества (вознаграждения) [15]. Цель агента – максимизировать математическое ожидание суммарного дисконтированного вознаграждения.

Среда описывается марковским процессом принятия решений, задаваемым пятеркой

$$M = (S,\, A,\, P,\, R,\, \gamma),$$

где S – множество состояний; A – множество действий; $P(s \mid s,\, a)$ – вероятностный закон переходов; $R(s,\, a)$ – вознаграждение; $\gamma \in (0,\, 1)$ – фактор дисконтирования.

Политика агента $\pi(a \mid s)$ задает условное распределение действий a при данном состоянии s. В условиях финансового рынка агент наблюдает не полное состояние, а лишь конечный набор признаков x_t. Для учета частичной наблюдаемости используется рекуррентное скрытое состояние h_t, аккумулирующее информацию о прошлых наблюдениях.

В работе рассматривается приближенная постановка задачи, в которой состоянием считается вектор

$$s_t = (x_t,\, w_{t-1},\, h_{t-1}) \in S.$$

Действием является вектор целевых весов

$$\hat{w}_t \in R^N,$$

генерируемый политикой агента [10, 17, 18, 26].

Фактические веса портфеля получаются проекцией на множество допустимых портфелей:

$$w_t = \Pi(\hat{w}_t),$$

где оператор $\Pi(\cdot)$ реализует приведение к ограничениям (1), (2); переход среды $s_t \to s_{t+1}$ определяется реализацией рыночных доходностей r_{t+1}, обновлением весов w_t, пересчетом признаков x_{t+1} и обновлением скрытого состояния h_t; вознаграждение R_t в момент t задается разработанной автором многокритериальной функцией вознаграждения:

$$R_t = r_{p,t+1} - \lambda_1 \rho_t^2 - \lambda_2 TC_t,$$

где первое слагаемое соответствует доходности портфеля после вычета издержек (4), второе — штрафу за сопряженность с бенчмарком (5), третье – штрафу за трансакционные издержки, зависящие от торгового оборота (3); коэффициент $\lambda_1 > 0$ задает относительный вес требований рыночной нейтральности, $\lambda_2 > 0$ – относительный вес требований к уровню издержек.

В экспериментальной постановке коэффициенты штрафов фиксировались на уровнях $\lambda_1 = 10$ для штрафа за корреляцию с бенчмарком и $\lambda_2 = 5$ для штрафа за трансакционные издержки.

Эти значения использовались во всех основных экспериментах и были выбраны по результатам кросс-валидации. Такой выбор коэффициентов обеспечивает баланс между снижением корреляции с бенчмарком и контролем трансакционных издержек при ребалансировке портфеля.

Цель агента – найти политику π, максимизирующую дисконтированную сумму вознаграждений:

$$J(\pi) = E_\pi\left[\sum_{t=0}^{T-1} \gamma^t R_t\right].$$

В предлагаемой методологии политика $\pi_\theta(a \mid s)$ параметризуется глубокой нейронной сетью. Обновление параметра θ выполняется с использованием алгоритма Proximal Policy Optimization (PPO), при этом рекуррентное скрытое состояние h_t реализуется на базе рекуррентного блока с вентильным механизмом Gated Recurrent Unit (GRU) и аппроксимирует скрытое состояние частично наблюдаемой среды.

2.3. Архитектура решения

В предлагаемой архитектуре используется иерархический экстрактор признаков на основе сверточных сетей (англ. *Convolutional Neural Network*, CNN) и рекуррентных блоков (GRU), поверх которого реализуется рекуррентный алгоритм

актор – критик (рис. 1). Целью разработки архитектуры является построение пространственно-временного представления состояния среды агента для оптимизации портфеля, которое одновременно учитывает различия между активами и динамику во времени.

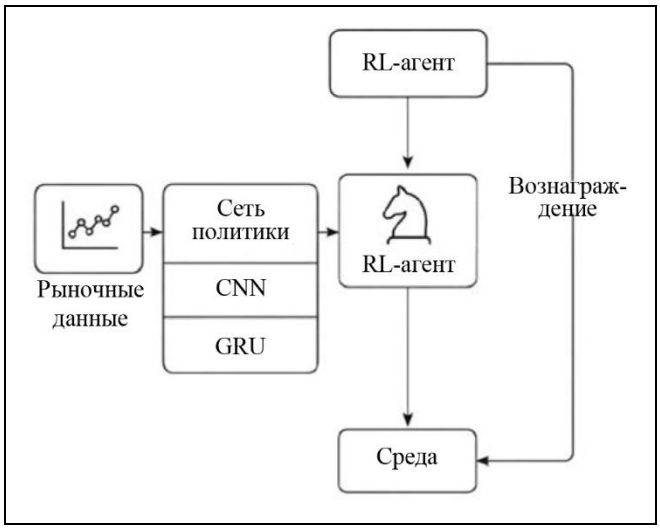

Рис. 1. Верхнеуровневая схема работы системы с интегрированной функцией вознаграждения для управления агентом

Пусть на шаге t на вход подается тензор признаков

$$X_t \in R^{N\times L\times F},$$

где N – число активов; L – число временны́х лагов; F – число признаков, характеризующих актив. Такой тензор интерпретируется как кросс-секционная карта признаков: по одной оси перечисляются активы, по второй – лаги, по третьей – признаковое описание.

Сверточные слои применяются к тензору X_t и реализуют локальную фильтрацию по оси актив – лаг, что позволяет выявлять пространственные и краткосрочные временны́е зависимости [10, 12, 26].

Результатом работы этого блока является тензор признаков

$$Z_t = CNN\left(X_t\right) \in R^{N\times F},$$

который далее агрегируется в вектор фиксированной размерности $z_t \in R^{d_z}$, где d_z – размерность агрегированного вектора признаков после сверточного блока.

Последовательность векторов $\{z_t\}$ подается в рекуррентный модуль GRU, реализующий эволюцию скрытого состояния:

$$h_t = GRU\left(z_t, h_{t-1}\right),\ h_t \in R^{d_h},$$

где d_h – размерность скрытого состояния рекуррентного блока GRU.

Скрытое состояние h_t аккумулирует информацию о прошлой динамике признаков и служит аппроксимацией состояния в частично наблюдаемой среде [17, 18, 26]. Поверх скрытого состояния строятся два выхода:

- Актор π_θ, отображающий состояние h_t (совместно с позиционной информацией, в том числе w_{t-1}) в вектор параметров действия $\hat{w}_t$, который затем проецируется на допустимое множество портфельных весов.
- Критик V_ϕ, отображающий состояние h_t в скалярную оценку ценности $V_\phi\left(h_t\right)$, используемую при вычислении преимуществ и стабилизации обучения [12, 13, 16, 19, 22].

Таким образом, реализуется рекуррентная архитектура актор – критик, в которой скрытое состояние h_t выступает общим представлением для политики и функции ценности.

Таким образом, комбинация нейронных сетей позволяет одновременно захватывать локальные пространственные закономерности и долгосрочну́ю временну́ю динамику, формируя расширенное представление состояния, на основании которого рекуррентный алгоритм актор – критик оптимизирует стратегию принятия инвестиционных решений [10, 17, 18].

На рис. 2 представлена схема архитектуры решения, где показаны основные блоки нейронных сетей и их интеграция в алгоритм.

3. ЭКСПЕРИМЕНТАЛЬНАЯ ВАЛИДАЦИЯ МЕТОДОЛОГИИ

3.1. Данные и методы

Для валидации методологии использовался панельный набор рыночных данных по компонентам фондовых индексов США, Европы и Азии. В выборку включены индексы: S&P 500 (GSPC), NASDAQ-100 (NDX), DJIA (DJI), FTSE 100 (FTSE), DAX (GDAXI), Hang Seng (HSI) и SSE Composite (SSEC). Данные загружены из терминала Bloomberg и охватывают период с января 2010 г. по декабрь 2020 г. Частота наблюдений – дневная. Для каждого индекса использовался полный состав компонентов, действующий на начало соответствующего обучающего окна; изменения состава учитывались автоматически в процессе

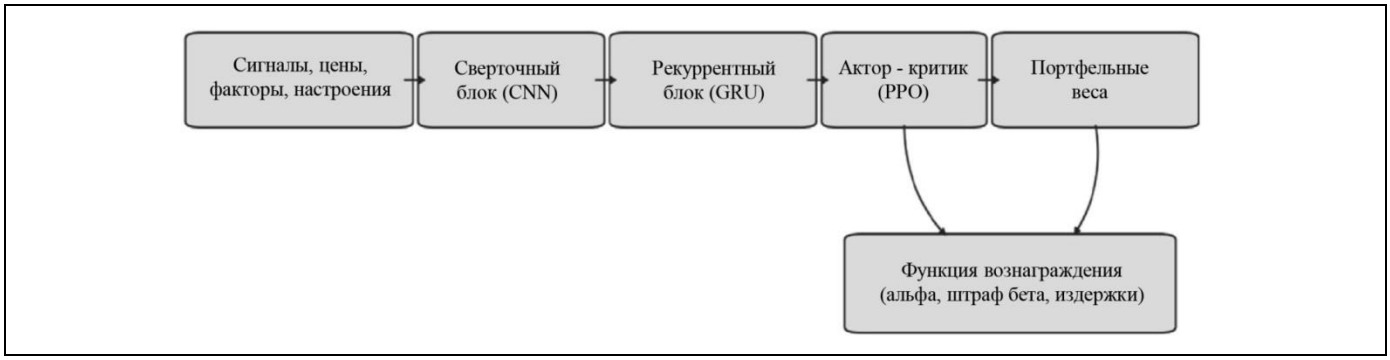

Рис. 2. Замкнутый контур обучения и принятия решений агентов в предлагаемой методологии

формирования панельного набора. Для каждого актива формировался вектор признаков, включающий следующие группы переменных: ценовые характеристики (логарифмические доходности за горизонты 1, 5, 10 и 20 дней); скользящие средние цен; реализованную волатильность; показатели ликвидности и объема: нормализованный объем торгов, спред, оборот; меры риска (коэффициенты вариации доходностей, индикаторы рыночной волатильности); макроэкономические и альтернативные признаки (процентные ставки, инфляционные показатели, индексы деловой активности, индикаторы настроений), т. е. стандартные показатели из терминала Bloomberg.

Для каждого момента времени t формировался тензор признаков

$$X_t \in R^{N_t \times L \times F},$$

где N_t – число активов; L – число временных лагов.

Предобработка выполнялась отдельно на каждом обучающем окне и включала следующие процедуры: синхронизацию данных по календарю и перенос последнего наблюдения в случае пропуска; винсоризацию по первому и 99-му перцентилям; стандартизацию признаков по обучающему сегменту; создание временных лагов показателей; формирование позиционных признаков (веса портфеля на предыдущем шаге, агрегированные показатели риска и ликвидности). Такая процедура исключает утечку информации и обеспечивает корректное разделение данных по временны́м сегментам.

Оценка обобщающей способности модели выполнялась по схеме пошаговой оптимизации. Каждый цикл включал обучающее окно длиной 750 дней, затем валидационное окно длиной 250 дней и далее тестовое окно длиной 250 дней. После завершения цикла окно сдвигалось на 250 дней, и процедура повторялась. На каждом цикле модель обучалась заново. На тестовом участке проводилась ежедневная ребалансировка портфеля с учетом трансакционных издержек и стоимости удержания коротких позиций. Метрики эффективности рассчитывались только на строго тестовых данных и усреднялись по всем тестовым окнам. Состояние среды включало тензор наблюдений, вектор весов портфеля и дополнительные агрегированные параметры (ликвидность, волатильность, скрытое состояние рекуррентного блока).

Псевдокод алгоритма приведен в Приложении.

3.2. Сравнительные модели и метрики

Было проведено сопоставление модели с перечнем референсных подходов для глубокого анализа рассматриваемой методологии:

- Mean–Variance – классическая оптимизация портфеля Марковица, где на каждом обучающем окне оцениваются ожидаемые доходности и ковариационная матрица, после чего строится портфель, минимизирующий дисперсию при заданной целевой доходности; в рамках валидации применяется с учетом реалистичных трансакционных издержек и дневной ребалансировки [5, 6, 9, 20].
- FF3 – факторная стратегия на базе модели Фамы – Френча, реализуемая через репликацию факторных портфелей; служит эталоном для измерения систематической факторной экспозиции портфеля [1, 2].
- Pairs Trading – парная стратегия, основанная на выборе коинтегрированных ликвидных пар из состава индекса; она запускается и используется на отклонениях спреда от равновесного значения. Такая стратегия обеспечивает локальную рыночно-нейтральную позицию, но остается чувствительной к ликвидности и параметрам отбора торговых пар [5, 8, 20].
- RL-Baseline – примитивный агент обучения с подкреплением, использующий те же входные признаки и схему валидации, что и продвинутый агент; служит контрольной точкой для оценки

преимуществ сложных планирующих глубоких методов продвинутого агента [10, 17, 21, 26].

• AlphaZeroBeta – предлагаемая модель, которая объединяет рекуррентную нейронную архитектуру и многокритериальную функцию вознаграждения, направленную на одновременное повышение риск-скорректированной доходности, снижение корреляции с рыночным индексом и ограничение оборота.

Используются три метрики:

• Коэффициент Шарпа (англ. *Sharpe ratio, Sharpe*) – характеризует риск-скорректированную отдачу, т. е. отношение средней избыточной доходности портфеля к волатильности этой доходности; более высокие значения указывают на более эффективное сочетание доходности и риска [28]:

$$Sharpe = \frac{E\left[r_{p,t+1} - r_f \right]}{\sqrt{\mathrm{Var}\left(r_{p,t+1} \right)}},$$

где r_f – безрисковая ставка (ставка, приведенная к дневному исчислению из годовой).

• Максимальная просадка (англ. *Max Drawdown*, MDD) – отражает наихудшее относительное падение накопленной стоимости портфеля от предшествующего пика до локального минимума за рассмотренный период и показывает уязвимость стратегии к крупным просадкам [13, 22]:

$$MDD = \max_{0<t<s<T} \frac{W_t - W_s}{W_t},$$

где W_t – накопленная стоимость портфеля в момент *t*, т. е. стоимость портфеля до начала падения; W_s – накопленная стоимость портфеля в более поздний момент *s*, после снижения.

• Корреляция с рыночным индексом характеризует линейную зависимость между доходностями портфеля и бенчмарка, служит проксимальной оценкой экспозиции к рыночной бете (положительная корреляция означает, что стоимость портфеля изменяется параллельно рынку) [2, 6]:

$$\mathrm{Corr} = \frac{\mathrm{Cov}\left(r_{p,t+1},\, m_{t+1} \right)}{\sqrt{\mathrm{Var}\left(r_{p,t+1} \right) \mathrm{Var}\left(m_{t+1} \right)}}.$$

3.3. Параметры модели, архитектуры и данных

В экспериментах использовались следующие фиксированные параметры:

• Для пошаговой оптимизации применялись окна длиной $T_{\text{train}} = 750$ дней (≈ 3 года) для обучения, $T_{\text{train}} = 250$ дней (≈ 1 год) для подбора гиперпараметров и ранней остановки и $T_{\text{test}} = 250$ дней для оценки метрик, с последующим сдвигом на T_{test}.

• Размер временного лага признаков $L = 20$ дней.

• Архитектура CNN включала два сверточных слоя с 32 и 64 фильтрами и ядром 3×3, за которыми следовали агрегация и полносвязный слой размерности 128. Рекуррентный модуль GRU имел скрытое состояние размерности $d_h = 128$.

• Актор представлялся двухслойной полносвязной сетью $128 \to 64 \to N$, критик – сетью $128 \to 64 \to 1$.

• Обучение выполнялось алгоритмом PPO со следующими параметрами: коэффициентом дисконтирования $\gamma = 0{,}99$, коэффициентом обобщенной оценки преимущества $\lambda_{\text{GAE}} = 0{,}95$, коэффициентом клиппинга $\varepsilon_{\text{PPO}} = 0{,}20$.

• Коэффициент регуляризации критика $c_{\text{v}} = 1{,}0$.

• Коэффициент энтропийной регуляризации $c_{\text{ent}} = 0{,}01$.

• Длина окна составляла 64 шагов, размер мини-батча – 256, количество эпох оптимизации PPO на один шаг обучения – 10.

• В функции вознаграждения использовались коэффициенты штрафов $\lambda_1 = 10$ (корреляция с бенчмарком), $\lambda_2 = 5$ (трансакционные издержки).

• Стоимость линейной компоненты издержек $c_{\text{lin}} = 1 \cdot 10^{-4}$, квадратичной $c_{\text{quad}} = 1 \cdot 10^{-6}$.

• Параметры ограничения портфеля: максимальная валовая экспозиция $\Gamma = 1{,}0$, ограничение изменения весов до 0,2 на одну инвестированную единицу капитала.

В качестве источника данных использовался терминал Bloomberg; для компонент индексов GSPC, NDX, DJI, FTSE, GDAXI, HSI, SHCOMP применялись ценовые ряды PX_LAST, PX_OPEN, PX_HIGH, PX_LOW, объемы VOLUME и, при наличии, показатели ликвидности (BID, ASK, BID_ASK_SPREAD). Для построения рыночных и макроэкономических факторов использовались: индикаторы волатильности VIX, V2X, HSIV; процентные ставки USGG2YR, USGG5YR, USGG10YR, US0003M; инфляционные и макроэкономические ряды CPI YOY, PPI YOY, GDP CQOQ, UNRATE, IP CHNG; индикаторы деловой активности NAPMPMI, NAPMNMI, ECOPMIM, CHPMMAN; прокси настроений и риск-аппетита NHBUS, CPSENT, BAMLELL,

NEWS_SENTIMENT; опционные показатели SDEX, SPX SKEW, GVZ, OVX; а также глобальные валютные и товарные индикаторы DXY, EURUSD, JPYUSD, CL1, GC1. Все данные загружались с дневной частотой за период 2010–2020 гг., синхронизировались по календарю и проходили процедуру нормализации по статистике обучающих окон.

3.4. Результаты и диагностика обучения

Результаты анализа показателей эффективности (табл. 1–3) демонстрируют, что рассматриваемый подход AlphaZeroBeta уверенно лидирует по коэффициенту Шарпа, сохраняет близкую к нулю корреляцию с рыночным индексом и заметно уменьшает максимальную просадку портфеля в периоды рыночных шоков. Такое сочетание высокой риск-скорректированной доходности и низкой систематической зависимости свидетельствует о наличии инвестиционного преимущества подхода, а также об эффективных механизмах управления рисками [5, 10, 13, 20, 22, 28].

Таблица 1

Значения коэффициента Шарпа для разных стратегий

Стратегия	США	Европа	Азия
Mean–Variance	0,90	0,80	0,70
FFE (Fama-French)	0,95	0,85	0,75
Pairs Trading	0,80	0,70	0,65
RL-Baseline	1,10	0,90	0,80
AlphaZeroBeta (предлагаемый подход)	1,35	1,20	1,05

Таблица 2

Корреляция стратегий с рыночными индексами

Стратегия	США	Европа	Азия
Mean–Variance	0,55	0,60	0,50
FFE (Fama-French)	0,45	0,50	0,40
Pairs Trading	0,35	0,30	0,25
RL-Baseline	0,20	0,15	0,10
AlphaZeroBeta (предлагаемый подход)	0,02	0,01	0,00

Таблица 3

Максимальная просадка стратегий, %

Стратегия	США	Европа	Азия
Mean–Variance	–35	–32	–30
FFE (Fama-French)	–28	–26	–24
Pairs Trading	–25	–22	–20
RL-Baseline	–18	–17	–15
AlphaZeroBeta (предлагаемый подход)	–12	–11	–10

Экспериментальные результаты показывают, что агент адаптивно оптимизирует распределение капитала, сохраняя эффективность в кризисные и волатильные периоды; включение в функцию вознаграждения штрафа за корреляцию с бенчмарком поддерживает близкую к нулю рыночную экспозицию без жестких ограничений, что снижает чувствительность к макроэкономическим шокам [1, 2, 13]. Учет оборота и трансакционных издержек непосредственно на этапе обучения повышает практическую реализуемость и уменьшает риск систематической переоценки результатов в тестировании [5, 13, 14].

В совокупности алгоритм демонстрирует улучшенные риск-скорректированные значения метрик, меньшие максимальные просадки портфеля по сравнению с альтернативными подходами. Ключевые ограничения связаны с калибровкой весов целевой функции, качеством данных и моделью издержек. Метод выделяется целевой оптимизацией нейтральности через компонент вознаграждения (перевод «нулевой беты» из побочного эффекта в целевую характеристику стратегии); гибридный нейросетевой экстрактор в связке с рекуррентной политикой повышает устойчивость к смене режимов и частичной наблюдаемости, а учет оборота приближает работу моделей к реальным торговым условиям [5, 16–18, 20, 22, 28].

Все вычисления проводятся на GPU-инфраструктуре класса V100/A100; один запуск обучается несколько часов, а полная серия из 20–30 скользящих окон на один индекс укладывается в 5–10 дней на единичном GPU (или в 1–3 дня при параллелизации запусков на 4–8 GPU).

На рис. 3 приведено сравнение коэффициентов Шарпа для пяти стратегий (Mean–Variance, FF3, Pairs Trading, RL-Baseline и AlphaZeroBeta) по трем регионам (США, Европа, Азия); для расчета коэффициента Шарпа использовалась отложенная

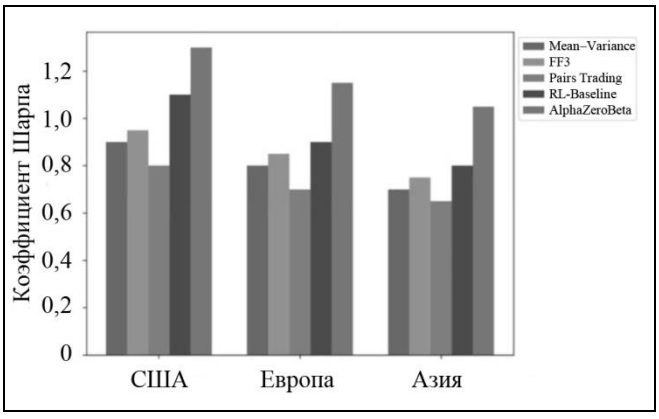

Рис. 3. Средние коэффициента Шарпа на отложенной выборке (OOS) для разных стратегий по регионам

выборка (англ. *out-of-sample,* OOS), что соответствует стандартной практике оценки эффективности алгоритма. Результаты показывают, что предлагаемый подход демонстрирует наивысшие значения коэффициента Шарпа по всем регионам, опережая и классические портфельные модели, и портфели однофакторных RL-агентов [5, 10, 20, 26, 28]. На рис. 4 показана динамика корреляции (рыночной беты) стратегий с соответствующими бенчмарками: подход поддерживает близкую к нулю или отрицательную корреляцию благодаря целевому штрафу за сопряженность с рынком в функции вознаграждения, тогда как референтные подходы (включая факторные и парные стратегии) сохраняют заметную положительную связь с индексом [1, 2, 6, 8, 13].

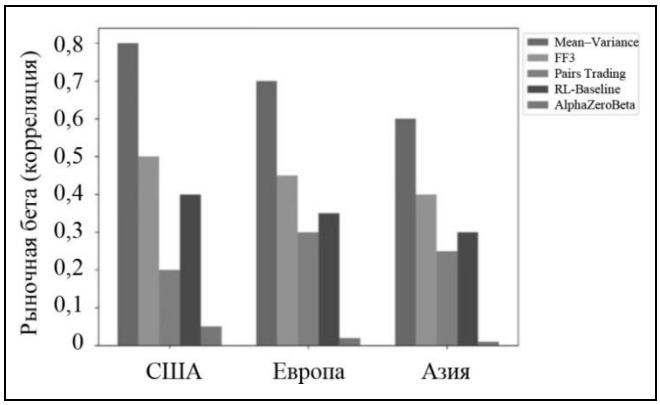

Рис. 4. Средние значения коэффициента корреляции Пирсона на отложенной выборке (OOS) с бенчмарками

Нормализация признаков и вычисление метрик выполнялись строго на тестовых сегментах с исключением любой информации из обучающих окон, чтобы избежать утечки данных [9, 11]. Транзакционные издержки моделировались и вычитались при формировании портфельных доходностей, что приближает результаты к реалистичному исполнению [5, 13, 14].

4. НАПРАВЛЕНИЯ ДАЛЬНЕЙШИХ ИССЛЕДОВАНИЙ

В качестве приоритетных направлений дальнейших исследований можно выделить следующие:

• Усилить интерпретируемость модели – применить методы анализа моделей машинного обучения, чтобы напрямую сопоставлять вклад сигналов стратегии с классическими факторами (FF3, Carhart) и демонстрировать экономическую интерпретируемость и атрибуцию выявленного инвестиционного преимущества; такие инструменты помогают отличить «черный ящик» от воспроизводимого экономического эффекта и связать инвестиционное преимущество с известными факторами фундаментальной экономической литературы.

• Адаптировать алгоритм для работы на кросс-активных торговых вселенных с учетом маржинальных требований и клиринговых ограничений, что потребует интеграции моделей ликвидности в торговую среду агента и стресс-тестирования под различными маржинальными шоками.

• Разработать и формализовать онлайн-обучение и стабилизирующие механизмы, например ограничение дивергенции при обновлении политики агента.

• Усилить устойчивость модели через симуляцию шоков ликвидности, искажение признаков, а также применение риск-чувствительных метрик (более продвинутые метрики риска) при оптимизации и оценке вознаграждения, чтобы контролировать экспозицию портфеля в стресс-сценариях.

• Провести систематическую валидацию по компонентам архитектуры, анализ чувствительности модели по параметрам функции вознаграждения и проскальзывания; задокументировать и сделать слепок данных и метрик для воспроизводимости.

• Объединить перечисленные техники с инфраструктурными требованиями, такими как задержки исполнения и реалистичные модели стоимости активов, чтобы обеспечить воспроизводимость результатов тестирования на исторических данных в реальных условиях.

ЗАКЛЮЧЕНИЕ

В работе представлена методология построения рыночно-нейтральных стратегий, формализующая распределение капитала как последовательный процесс принятия решений с учетом трансакционных издержек и прямого контроля корреляции с рыночным индексом. Включение реалистичной модели издержек и эксплуатационных ограничений снижает риск систематической переоценки результатов и повышает практическую пользу подхода путем учета издержек на этапе обучения. Многоцелевой функционал вознаграждения, который балансирует доходность, штрафы за экспозицию к рынку и торговый оборот, объединен с рекуррентной архитектурой актор – критик. Такое сочетание обеспечивает адаптивность к нестационарности, устойчивость к смене рыночных режимов и снижение дисперсии градиентных оценок. Статические факторные методы проще с точки зрения реализации и интерпретируемости, но уступают предлагаемому подходу по адаптивности

и учету эксплуатационных ограничений. Для контроля рисков в дальнейшем важно учитывать долю риска в целевой функции, что повысит экономическую интерпретируемость и устойчивость результатов в стресс-сценариях.

Благодарности. *Автор выражает благодарность коллегам из Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ) за ценные обсуждения, конструктивные отзывы и постоянную поддержку при подготовке этой работы.*

ПРИЛОЖЕНИЕ

Псевдокод алгоритма

Алгоритм 1. Обучение и тестирование рыночно-нейтральной стратегии
Входные данные: D – панель рыночных данных, M – доходности бенчмарка, U – множество активов, H – гиперпараметры модели, обучения и ограничений портфеля.
Результат: лучшая обученная политика для каждого окна, метрики вне выборки по всем окнам

```
S ← Построить последовательность скользящих окон (обучение, валидация, тест)
Для каждого окна s ∈ S:
    (X_train, R_train, M_train, stats) ← Предобработка(D_train, M_train, только по обучающему окну)
    (X_val,   R_val,   M_val)          ← Предобработка(D_val,   M_val,   stats)
    (X_test,  R_test,  M_test)         ← Предобработка(D_test,  M_test,  stats)
Инициализировать параметры кодировщика, актора и критика
best_score ← −∞
Пока не выполнено условие ранней остановки:
    Очистить буфер траекторий
    h ← 0; w_prev ← 0
    Инициализировать состояние онлайн-оценки корреляции
    Для каждого момента t на обучающем окне:
        h_t ← Кодировщик(X_train[t], h_{t−1})
        a_t ← Актор(h_t)
        w*_t ← Преобразовать действие в целевые веса
        w_t  ← Проецировать на допустимое множество
        w_pre ← Обновить веса после рыночного движения
        Δw_t  ← Ограничить изменение весов по обороту
        TC_t  ← Рассчитать трансакционные издержки(Δw_t)
        w_t   ← Применить исполнение сделки
        p_t^gross ← Доходность портфеля за следующий шаг до вычета издержек
        ρ_t ← Обновить онлайн-оценку корреляции(p_t^gross, M_train[t+1])
        r_t ← p_t^gross − λ_corr · ρ_t^2 − λ_tc · TC_t
        V_t ← Критик(h_t)
        Сохранить (h_t, a_t, w_t, p_t^gross, ρ_t, TC_t, r_t, V_t) в буфер
        w_prev ← w_t
    Оценить преимущества и целевые значения критика по буферу
    Обновить параметры модели методом PPO на усечённых последовательностях
    score_val ← Проверить политику на валидационном окне по выбранному критерию
    Если score_val лучше best_score:
        Сохранить текущие параметры как лучшие
        best_score ← score_val
    Загрузить лучшую модель для данного окна
    test_metrics ← Протестировать политику на тестовом окне
    Сохранить test_metrics
Объединить метрики по всем окнам
Вернуть агрегированные результаты и результаты по окнам
```

ЛИТЕРАТУРА

1. *Carhart, M.M.* On Persistence in Mutual Fund Performance // The Journal of Finance. – 1997. – Vol. 52, no. 1. – P. 57–82.
2. *Fama, E.F., French, K.R.* Common Risk Factors in the Returns on Stocks and Bonds // Journal of Financial Economics. – 1993. – Vol. 33, no. 1. – P. 3–56.
3. *Goyal, A., Welch, I.* A Comprehensive Look at the Empirical Performance of Equity Premium Prediction // The Review of Financial Studies. – 2008. – Vol. 21, no. 4. – P. 1455–1508.
4. *Lettau, M., Ludvigson, S.* Consumption, Aggregate Wealth, and Expected Stock Returns // The Journal of Finance. – 2001. – Vol. 56, no. 3. – P. 815–849.
5. *Pedersen, L.H.* Efficiently Inefficient: How Smart Money Invests and Market Prices Are Determined. – Princeton: Princeton University Press, 2015. – 368 p.
6. *Markowitz, H.* Portfolio Selection // The Journal of Finance. – 1952. – Vol. 7, no. 1. – P. 77–91.
7. *Loomis, C.J.* The Jones Nobody Keeps up with // Fortune Magazine. – April 1, 1966. – P. 237–247.
8. *Gatev, E., Goetzmann, W.N., Rouwenhorst, K.G.* Pairs Trading: Performance of a Relative-Value Arbitrage Rule // The Review of Financial Studies. – 2006. – Vol. 19, no. 3. – P. 797–827.
9. *Choi, I.* Efficient Estimation of Factor Models // Econometric Theory. – 2012. – Vol. 28, no. 2. – P. 274–308.
10. *Jiang, R., Xu, D., Liang, Z.* A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem // arXiv:1706.10059. – 2017. – DOI: https://doi.org/10.48550/arXiv.1706.10059
11. *Gu, S., Kelly, B., Xiu, D.* Empirical Asset Pricing via Machine Learning // The Review of Financial Studies. – 2020. – Vol. 33, no. 5. – P. 2223–2273.
12. *Srivastava, N., Hinton, G., Krizhevsky, A.,* et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research. – 2014. – Vol. 15, no. 56. – P. 1929–1958.
13. *Buehler, H., Gonon, L., Teichmann, J., Wood, B.* Deep Hedging // Quantitative Finance. – 2019. – Vol. 19, no. 8. – P. 1271–1291.
14. *Moody, J., Saffell, M.* Learning to Trade via Direct Reinforcement // IEEE Transactions on Neural Networks. – 2001. – Vol. 12, no. 4. – P. 875–889.
15. *Sutton, R.S., Barto, A.G.* Reinforcement Learning: An Introduction. – 2nd ed. – Cambridge, MA: MIT Press, 2018. – 552 p.
16. *Schulman, J., Wolski, F., Dhariwal, P.,* et al. Proximal Policy Optimization Algorithms // arXiv:1707.06347. – 2017. – DOI: https://doi.org/10.48550/arXiv.1707.06347
17. *Heess, N., Wayne, G., Silver, D.,* et al. Memory-Based Control with Recurrent Neural Networks // arXiv:1512.04455. – 2015. – DOI: https://doi.org/10.48550/arXiv.1512.04455
18. *Hausknecht, M., Stone, P.* Deep Recurrent Q-Learning for Partially Observable MDPs // arXiv:1507.06527. – 2015. – DOI: https://doi.org/10.48550/arXiv.1507.06527
19. *Silver, D., Lever, G., Heess, N.,* et al. Deterministic Policy Gradient Algorithms // Proceedings of the 31st International Conference on Machine Learning (ICML 2014). – Beijing, 2014. – Vol. 32, no. 1. – P. 387–395.
20. *Bali, T.G., Brown, S.J., Caglayan, M.O.* Macroeconomic Risk and Hedge Fund Returns // Journal of Financial Economics. – 2014. – Vol. 114, no. 1. – P. 1–19.
21. *Kogan, L., Papanikolaou, D.* Firm Characteristics and Stock Returns: The Role of Investment-Specific Shocks // The Review of Financial Studies. – 2013. – Vol. 26, no. 11. – P. 2718–2759.
22. *Rockafellar, R.T., Uryasev, S.* Optimization of Conditional Value-at-Risk // Journal of Risk. – 2000. – Vol. 2, no. 3. – P. 21–41.
23. *Liu, X., Xiong, Z., Zhong, S.,* et al. Practical Deep Reinforcement Learning Approach for Stock Trading // arXiv:1811.07522. – 2018. – DOI: https://doi.org/10.48550/arXiv.1811.07522
24. *Liu, X., Yang, H., Gao, J., Wang, C.D.* FinRL: Deep Reinforcement Learning Framework to Automate Trading in Quantitative Finance // arXiv:2111.09395. – 2021. – DOI: https://doi.org/10.48550/arXiv.2111.09395
25. *Gu, F., Jiang, Z., García-Fernández, Á.F.,* et al. MTS: A Deep Reinforcement Learning Portfolio Management Framework with Time-Awareness and Short-Selling // arXiv:2503.04143. – 2025. – DOI: https://doi.org/10.48550/arXiv.2503.04143
26. *Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.* Deep Reinforcement Learning: A Brief Survey // IEEE Signal Processing Magazine. – 2017. – Vol. 34, no. 6. – P. 26–38.
27. *Choudhary, H., Orra, A., Sahoo, K., Thakur, M.* Risk-Adjusted Deep Reinforcement Learning for Portfolio Optimization: A Multi-reward Approach // International Journal of Computational Intelligence Systems. – 2025. – Vol. 18, no. 1. – P. 1–19.
28. *Sharpe, W.F.* Mutual Fund Performance // The Journal of Business. – 1966. – Vol. 39, no. 1. – P. 119–138.

Статья представлена к публикации членом редколлегии В. В. Клочковым.

Поступила в редакцию 03.10.2025,
после доработки 11.03.2026.
Принята к публикации 08.04.2026.

Беляков Борис Евгеньевич – аспирант, Московский институт электроники и математики им. А.Н. Тихонова НИУ ВШЭ; приглашенный преподаватель, Национальный исследовательский университет «Высшая школа экономики», г. Москва
✉ work.belyakov@gmail.com,
ORCID iD: https://orcid.org/0009-0007-8237-9833

APPLICATION OF A MULTICRITERIA REWARD FUNCTION TO DYNAMIC RISK MANAGEMENT IN MARKET-NEUTRAL PORTFOLIOS

B. E. Belyakov

HSE Tikhonov Moscow Institute of Electronics and Mathematics (HSE MIEM), Moscow, Russia
National Research University Higher School of Economics (HSE), Moscow, Russia

✉ work.belyakov@gmail.com

Abstract. In this paper, a multicriteria reward function is developed and empirically validated within a market-neutral portfolio management framework based on deep reinforcement learning. This function formalizes the optimal capital allocation problem as a Markov decision process in a nonstationary environment. The reward function simultaneously optimizes excess return per unit of risk, market neutrality via an explicit penalty on the correlation with a target market index, and transaction costs. As a result, systemic risk control and the cost of portfolio rebalancing are embedded into the modeling and training procedure. The model is trained and analyzed using cross-validation to mitigate the bias and adapt to regime shifts. According to the experimental validation results on leading equity indices of the USA, Europe, and Asia over a ten-year horizon, the above framework is effective, particularly showing consistent improvements in key metrics: higher Sharpe ratios, lower maximum drawdowns, and a substantially reduced correlation with the market index relative to classical optimization algorithms and reinforcement learning approaches with a single criterion.

Keywords: market-neutral strategy, portfolio optimization, reinforcement learning, market-neutral portfolio, deep learning, portfolio management.

Acknowledgments. The author is grateful to HSE colleagues for their valuable discussions, constructive feedback, and continued support of this work.