Обучение с подкреплением на основе моделей для оптимизации биржевой торговли

Хуэйфан Хуан, Тинг Гао, Луксуань Ян, И Гуй, Цзинь Го, Пэн Чжан

Сентябрь 2021 года

Аннотация

Обучение с подкреплением (RL) привлекает внимание все большего числа исследователей в области количественных финансов, поскольку структура взаимодействия агента и среды связана с процессом принятия решений во многих бизнес-задачах. Большинство современных финансовых приложений, использующих алгоритмы RL, основаны на методе без моделей, который по-прежнему сталкивается с проблемами стабильности и адаптивности. Поскольку многие передовые алгоритмы обучения с подкреплением на основе моделей (MBRL) уже используются в таких приложениях, как видеоигры или робототехника, мы разработали новый подход, который использует уровень сопротивления и поддержки (RS) в качестве условий регуляризации для действий в MBRL, чтобы улучшить эффективность и стабильность алгоритма. Из результатов экспериментов видно, что уровень RS, как техника рыночного тайминга, повышает производительность чистых моделей MBRL по различным измерениям и позволяет получить большую прибыль при меньшем риске. Кроме того, предложенный нами метод даже противостоит большому падению (меньше максимальной просадки) в период пандемии COVID-19, когда на финансовом рынке разразился непредсказуемый кризис. Объяснение того, почему контроль уровня сопротивления и поддержки может повысить MBRL, также исследуется с помощью численных экспериментов, таких как потери сети актор-критик и ошибка предсказания переходной динамической модели. Показано, что RS-индикаторы действительно помогают алгоритмам MBRL быстрее сходиться на ранних этапах и получать меньшие потери критики по мере увеличения количества обучающих эпизодов.

1 Введение

С развитием современных методов глубокого обучения все больше и больше передовых нейросетевых структур находят широкое применение в

различных областях исследований и приложений. Эти механизмы машинного обучения в свою очередь вдохновляют исследователей из разных областей на разработку более совершенных архитектур нейронных сетей для повышения производительности моделей и решения более сложных сценариев. Одна из перспективных областей применения - количественные финансы - в последнее время использует преимущества методов ИИ для создания множества инновационных алгоритмических торговых стратегий для финансовых инвестиций и оптимизации портфеля.

Среди всех фреймворков машинного обучения, обучение с подкреплением (RL) обладает уникальным преимуществом: интерактивный процесс обучения соответствует человеческому

принятия решений. За последние годы для решения многих задач РЛ были разработаны два основных способа оптимизации, Q-learning и Actor-Critic, а также их комбинация. Например, если взять финансовые приложения, то Пастор и другие [Pastore et al.(2016)] проанализировали данные о 46 игроках из онлайн-игр на финансовых рынках и проверили, может ли Q-Learning отразить поведение этих игроков на основе меры рискованности. Их результаты показывают, что не все игроки близоруки, что противоречит гипотезе наивного инвестора. Кроме того, Ли и др. [Li et al.(2019b)] изучают преимущества трех различных классических глубоких RL-моделей DQN [Mnih et al.(2013)], Double DQN [Van Hasselt et al.(2016)] и Dueling DQN [Wang et al.(2016)] при прогнозировании цены акций и приходят к выводу, что DQN имеет наилучшую производительность. Кроме того, некоторые исследователи моделируют и сравнивают улучшенный метод глубокого RL с алгоритмом Adaboost и предлагают гибридное решение. Интересно, что Ли и др. (2019) [Lee et al.(2019)] применяют CNN к глубокой Q-сети, которая принимает на вход цены акций и изображения графиков объемов для прогнозирования мировых фондовых рынков. Что касается градиента политики, то Канг и другие (2018) [Kang et al.(2018)] используют современный алгоритм Asynchronous Advantage Actor-Critic (A3C [Mnih et al.(2016)]) для решения задачи управления портфелем и разрабатывают автономную глубокую RL-модель. Кроме того, Ли и др. [Li et al.(2019a)] предлагают адаптивный глубокий детерминированный метод РЛ (Adaptive DDPG) для некоторой задачи распределения портфеля, который включает оптимистичный или пессимистичный глубокий алгоритм РЛ, основанный на влиянии ошибки прогнозирования. Проанализировав данные по 30 компонентам акций Dow Jones, торговая стратегия превосходит традиционный метод DDPG [Lillicrap et al.(2015)]. Сутта [Sornmayura(2019)] сравнивает производительность агента ИИ с результатами стратегии buy-and-hold и эксперта-трейдера, тестируя 15-летний рынок форекс с помощью парного t-Test. Результаты показывают, что ИИ может превзойти стратегию "купи и держи" и товарного торгового советника на FOREX для валютных пар EURUSD и USDJPY.

Тем не менее, все еще существуют сложные и трудные вопросы, связанные с обучением с подкреплением для финансовой среды. Прежде всего, данные финансового маркетинга в большинстве сценариев имеют сложную нелинейность, даже сложное хаотическое поведение и неопределенность, включая негауссовский шум, что приводит к смещению распределения временных рядов данных [Cai and Wei(2020)] с течением времени. Кроме того, скрытые взаимодействия между различными агентствами могут быть непредсказуемо сложными для понимания. Как показывают многочисленные маркетинговые анализы акций, агентства с большими капиталовложениями иногда могут вызывать резкие колебания цен, что приводит к маркетинговой панике и нестабильности среди населения. Поэтому ученые активно ищут внутренние механики для решения вышеуказанных проблем. С одной стороны, для предварительной обработки данных финансовых временных рядов используются различные методы обесцвечивания данных. Например, Бао В. и др.

(2017) [Bao et al.(2017)] предлагают новую модель для прогнозирования акций, вейвлет-преобразования для разложения и устранения шума временного ряда цен на акции, что помогает стековым автокодировщикам (SAE) для извлечения высокоуровневых глубоких признаков и долговременной памяти (LSTM) лучше прогнозировать цену акций и повышать точность прогнозирования и прибыльность. С другой стороны, исследователи также создают множество видов индексов для оптимизации прибыли портфеля. В общем, все это затрудняет построение

система обучения с подкреплением, которая напрямую взаимодействует со сложной реальной маркетинговой средой, и это вдохновило нас на исследование RL на основе моделей для алгоритмической оптимизации торговых стратегий.

Большая часть существующей литературы основана на методе безмодельного обучения с подкреплением (MFRL). Однако эти алгоритмы достаточно дороги в обучении, учитывая высокий риск потери инвестиций на ранних стадиях в условиях реального финансового рынка, даже при использовании симулированных доменов (Mnih et al., 2015 [Mnih et al.(2015)], Lillicrap et al., 2015 [Lillicrap et al.(2015)], Schulman et al., 2017 [Schulman et al.(2017)]). Перспективным направлением повышения эффективности выборки является изучение методов обучения с подкреплением на основе моделей (MBRL) [Yu et al.(2019)]. Dy- naQ [Peng et al.(2018)], PETS [Chua et al.(2018)], MBPO [Janner et al.(2019)] - мощные модельные алгоритмы подкрепления в среде Gym. Алгоритмы RL на основе моделей могут достичь превосходной эффективности выборки, но часто отстают от лучших алгоритмов без моделей в плане асимптотической производительности. Чуа, Калан-дра и др. (2019) [Chua et al.(2018)] изучают, как преодолеть этот разрыв, используя динамические модели, учитывающие неопределенность. Они предлагают новый алгоритм под названием PETS, который приближается к асимптотической производительности нескольких эталонных алгоритмов без моделей, требуя при этом значительно меньшего количества образцов. Яннер и др. (2019) [Janner et al.(2019)] предлагают алгоритм оптимизации политики на основе модели (MBPO) для повышения эффективности PETS за счет монотонного улучшения на каждом шаге и получения самой современной производительности. Все эти прекрасные работы вдохновляют нас на применение алгоритмов RL на основе моделей в сложной среде финансового маркетинга, и, насколько нам известно, в литературе редко встречаются работы по использованию MBRL в количественном финансовом моделировании.

Количественные финансы - это междисциплинарный предмет, в котором участвуют многие специалисты из разных областей с различными знаниями, и многие перспективные алгоритмические торговые стратегии в реальных задачах используют некоторые финансовые методы анализа. Например, для поиска наилучшего рыночного тайминга, уровень сопротивления (цена, при которой, по мнению трейдера, сила продавца начинает преобладать над силой покупателя) и уровень поддержки (цена, при которой, по мнению трейдера, сила покупателя начинает преобладать над силой продавца) часто рассматриваются как индикаторы рыночного тайминга для продажи и покупки соответственно. Поэтому мы используем возможности выбора времени по относительной силе сопротивления-поддержки (RSRS) и применяем их в классических алгоритмах MBRL, стремясь найти лучшие оптимизированные политики, которые могут улучшить алгоритмы MBRL, сделав их более стабильными, адаптивными и эффективными.

В целом, в данной работе сделан следующий основной вклад:

- **Брак RSRS и MBRL**. Мы разрабатываем финансовую среду для классических алгоритмов обучения с подкреплением на основе моделей (MBRL) и встраиваем относительную силу поддержки сопротивления (RSRS) в структуру RL для улучшения стабильности и адаптивности модели.
- **Эффективность бизнеса**. Относительная сила сопротивления-поддержки (RSRS), как мощная техника выбора времени, будучи встроенной в пару алгоритмов MBRL, помогает чистой стратегии MBRL получать высокие улучшения...

мента в семи различных измерениях, как с точки зрения управления рисками, так и с точки зрения оптимизации портфеля.

- **Производительность модели**. Чтобы лучше объяснить различия в производительности между MBRL с RSRS и без него в качестве регуляризации действий, мы также сравнили потери сети "актер-критик" среди всех алгоритмов. Из графиков потерь критики видно, что RSRS действительно помогает выбранным алгоритмам MBRL быстрее сходиться на ранних стадиях и получать меньшую ошибку, когда модели сходятся, что повышает эффективность модели.

- **Производительность переходной динамики**. Существует тесная связь между SAC для оптимизации политики при моделировании переходной динамики и стохастическим оптимальным управлением со стохастическими дифференциальными уравнениями в качестве ограничений на состояние. Мы также проверяем ошибку предсказания динамической модели перехода в терминах различных проекций координат, учитывая, что наше пространство состояний является сложным и высокоразмерным.

Мы строим нашу работу следующим образом: Сначала, в разделе 2.1, мы вводим некоторые исходные знания о RL на основе моделей, а затем объясняем нашу финансовую среду в рамках RL на основе моделей в разделе
2.2 и 2.3. Мы также интерпретируем динамику перехода с точки зрения динамической системы и описываем гауссовский процесс как некоторое стохастическое дифференциальное уравнение (SDE), и, соответственно, оптимизация политики SAC связана со стохастическим оптимальным управлением при ограничении SDE в разделе 2.4. Далее, в разделе 3 мы подробно объясняем алгоритм RSRS и представляем псевдокод нашей предложенной модели. Кроме того, в разделе 4 приведены все результаты экспериментов, а также подробные объяснения того, как и почему предложенный нами алгоритм лучше, чем чистый MBRL без RSRS. Наконец, в разделе 5 мы подводим итог нашим выводам и описываем направления будущих исследований.

2 Рамка

2.1 Общие сведения: RL на основе моделей

Обучение с подкреплением направлено на выработку оптимальной политики, максимизирующей ожидание кумулятивного вознаграждения в процессе взаимодействия с окружающей средой. В основном это методы без модели и с моделью, в зависимости от того, взаимодействует ли агент с чисто реальной средой.

Методы без моделей могут использоваться для решения многих сложных задач с наилучшей прогрессивной производительностью. Но обычно это требует большого количества взаимодействий с окружающей

средой и предъявляет относительно высокие требования к вычислительной мощности. Методы же, основанные на моделях, сосредоточены на построении модели среды, которую мы называем динамической моделью перехода, с высокой эффективностью выборки. Одна из целей обычно заключается в повышении точности оценок вероятности перехода из одного состояния в другое. Более того, когда изученная модель переходов близка к реальной среде, оптимальная стратегия может быть найдена непосредственно с помощью некоторого планирования

Традиционный процесс планирования включает в себя две части: поиск пути и оптимизацию траектории.

Если говорить более конкретно, то для поиска пути, когда пространство действий непрерывно, обычно используется CEM (Cross-Entropy Method). Это метод Монте-Карло, используемый в основном для оптимизации и выборки важности. Если пространство действий дискретно, то при большом пространстве поиска используется MCTS (Monte Carlo Tree Search). Оптимизация траектории относится к задаче решения целевой функции с ограничениями, то есть к задаче оптимального управления. Траекторная оптимизация обычно включает в себя методы стрельбы и коллокации. Методы стрельбы больше подходят для простых задач управления и задач без ограничений на траекторию, в то время как методы коллокации больше подходят для сложных задач управления и задач с ограничениями на траекторию.

В PETS параметрические модели подгоняются нейронной сетью, и предлагается алгоритм, объединяющий модель PE (Probabilistic Ensembles) и метод планирования TS (Trajectory Sampling). С помощью метода перекрестной энтропии (CEM) они указывают лучшее направление для действий, а затем делают выборку на его основе. В связи с определенным отклонением между установленной моделью окружающей среды и реальной моделью Яннер и Фу и др. (2019) [Janner et al.(2019)] предложили новый основанный на модели алгоритм MBPO для улучшения PET. Сформулировав и проанализировав основанный на модели RL с монотонным улучшением на каждом шаге, они изучают использование модели в оптимизации политики как теоретически, так и эмпирически. Затем они демонстрируют, что для оптимизации используется простая процедура планирования короткого развертывания на основе реальных данных и SAC (soft actor-critic) [Haarnoja et al.(2018)].

В сложных и шумных средах ошибка модели перехода может быть большой. Для обучения хорошей политике требуется точная модель, а для получения точной модели необходимо много взаимодействий с реальной средой. Пан и Хе (2020) предложили метод M2AC (Masked Model based Actor-Critic) [Pan et al.(2020)], который уменьшает влияние ошибки модели с помощью механизма маскировки и снимает проблемы переборчивости MBRL в случае малого объема данных и большого шума в реальной среде. Теоретически доказано, что разрыв между реальной доходностью и значением разворота маскированной ансамблевой модели может быть ограничен, если используемая индивидуальная модель имеет небольшую ошибку.

2.2 Определение пространства и вознаграждения

Рассматривая наше агентство (инвестора) как интеллектуального агента, а финансовый рынок - как соответствующую среду, мы моделируем проблему торговли акциями как марковский процесс принятия решений (MDP).

• **Пространство состояний** $\mathbb{S} = [B, P, W, I]$: Пространство состояний $\mathbb{S}$ - это собранная информация о рынке. $\forall s_t \in \mathbb{S}$ - это состояние агента в момент времени t, которое включает в себя информацию из четырех частей: баланс счета агентства B_t , текущая цена акций P_t , накопленная сумма.

холдинг W_t , технические индикаторы I_t . Здесь технические индикаторы I_t состоят из семи общих технологических факторов: MACD [Appel(2003)], SMA30, SMA60, BOLL [Bollinger(1992)], RSI [Țăran-Moroşan(2011)], CCI [Lai et al.(2020)], ADX. В таблице 1 подробно описаны обозначения состояний s_t .

Мы предполагаем, что эксперимент состоит из D акций.

Таблица 1: Условные обозначения государства

Условные обозначения	Определение
B_t	Остаток на счете в момент времени t; $B_t \in \mathrm{R}_+$
P_t	дневная цена закрытия каждой акции; $P_t \in \mathrm{R}_+^D$
W_t	совокупное владение акциями каждой из них; $W_t \in Z_+^D$
MACD	Дивергенция конвергенции скользящих средних: индикатор импульса отображает тренд
SMA30	30-дневная простая скользящая средняя: 30-дневная цена закрытия равна средневзвешенной
SMA60	60-дневная простая скользящая средняя: 60-дневная цена закрытия равна средневзвешенной
БОЛЛ	Полосы Боллинджера: судит о среднесрочном и долгосрочном тренде движения
RSI	Индекс относительной силы: определяет точки перегиба тренда
CCI	Индекс товарного канала: помогает определить степень отклонения цены
ADX	Средний индекс направленности: определяет силу тренда

• **Пространство действий** A: пространство действий A - это множество доступных операций во время транзакции над всеми запасами D. $\forall a_t \in \mathrm{A}$ - это действия, предпринимаемые агентом в момент времени t, предполагается, что они имеют конечную размерность и непрерывны. Здесь $_t$ - это D-мерное множество.
вектор, где размерность i_{th} представляет действие, совершенное над акцией i_{th}.

Если обозначить W^i как совокупное количество акций i_{th} в момент времени t, то доступные действия с акциями включают покупку, удержание и продажу. Мы предполагаем, что максимальный объем торгов для одной акции на одном шаге составляет 100 акций. Детали следующие:
• Покупка: $a_t^i = +h$, h акций могут быть куплены, и это приводит к $W_{t+1}^i = W_t^i + h$.
• Продажа: $a_t^i = -h$, h акций могут быть проданы из текущих запасов. В этом случае, $W_{t+1}^i = W_t^i - h$.
• Удержание: $a_t = 0$, что означает отсутствие изменений в W_t^i.
где $h \in [0, 100]$ - целое положительное число.

• **Вознаграждение** r_t : Функция вознаграждения - это отображение $\mathrm{R} : \mathrm{S} \times \mathrm{A} \to \mathrm{R}$. Как видно из дальнейшего уравнения (3), прямое вознаграждение r_t от выполнения действия a_t в состоянии s_t определяется как процент изменения суммы активов:

1) Сумма активов агента - это сумма оставшихся инвестиционных средств и текущей стоимости акций, которыми он владеет. Заметим, что

сумма активов агента в момент времени t - это $Asset_t$, а формула расчета - уравнение:

$$Актив_t = (B_t + P_t^T \cdot W_t) = B_0 - \sum_{\tau=0}^{t} P_\tau^T \cdot a_\tau + P_{t+1}^T \sum_{\tau=0}^{t} a_\tau \qquad (1)$$

2) Рассмотрим стоимость перехода, запишем ее как C_t :

$$C_t = P_t^T \cdot |a_t| \cdot процент_{затрат.} \qquad (2)$$

Абсолютное значение a_t здесь означает взятие абсолютного значения каждой компоненты a без изменения размерности вектора a_t .

3) Формула расчета прямого вознаграждения r_t такова:

$$r_t = \frac{Актив_{t+1} - Актив_t - C_t}{Активы_t} \times 100 \quad (3)$$

где B_t , P_t представлены в таблице 1; a_t - действие, выполняемое агентом в момент времени t; $P^T - W_t$ внутреннее произведение векторов.

В задаче RL целью агента является максимизация ожидаемого вознаграждения. Вознаграждение может оценить качество каждого действия. Более того, агент оптимизирует стратегию под руководством вознаграждения. Отмечая кумулятивное вознаграждение как R_t в момент времени t:

$$R_t = \Sigma_{\tau=1}^{t} r_\tau \quad (4)$$

2.3 Переходная динамика

Одним из критериев классификации алгоритмов RL является наличие динамической модели среды. RL на основе модели привлекателен тем, что динамическая модель не зависит от вознаграждения и может легко воспользоваться всеми достижениями глубокого контролируемого обучения для использования моделей с высокой производительностью [Chua et al.(2018)]. Однако недостатком этого метода является то, что во многих задачах обучения агенту зачастую сложно получить точную модель реального окружения, что приводит к большим накопленным ошибкам при взаимодействии с виртуальной средой. PETS, MBPO и M2AC - надежные алгоритмы, позволяющие сделать шаг к сокращению разрыва между методами RL, основанными и не основанными на моделях. Поэтому мы применяем эти модели на финансовом рынке для оптимизации биржевой торговли. Начнем с определения вероятности перехода.

• **Вероятность перехода** P: Изменение рыночных условий абстрагируется как функция перехода состояния. $P : S \times A \times S \rightarrow [0, 1]$ - это функция вероятностей переходов состояний.

$$P_{c,c}^{a} = P(s_{t+1} = s' \mid s_t = s, a_t = a)$$

.

В PETS, MBPO и M2AC для прогнозирования используется ансамблевый гауссовский процесс. динамика перехода состояний. Из стохастического анализа мы знаем, что состояние s_t удовлетворяет следующему стохастическому дифференциальному уравнению (СДУ):

$$dX_t = b(X_t, u_t)d_t + \sigma(X_t, u_t)dw_t \quad (5)$$

где начальное условие $X_0 = x \in \Omega$, $a_t \subset R^{da}$ - F_t -адаптированное управляющее поле, а w_t - d_w -мерное F_t -стандартное броуновское движение.

• **Политика** π: Политика - это отображение, характеризуемое политикой $\pi : S \rightarrow A$.

$\forall t \in 1, ..., T$. Это последовательность действий, заданная агентом, цель политики π агент должен максимизировать конечную ожидаемую стоимость портфеля.

Агент в состоянии $s_t \in \mathsf{S}$ выполняет действие $a_t \in \mathsf{A}$, следуя политике π, получает вознаграждение $r_t = \mathsf{R}(s_t, a_t)$ и переходит в следующее состояние s_{t+1} в соответствии с вероятностью перехода P. Агент взаимодействует с окружением для сбора траектории, а затем обновляет стратегию. Основная задача оптимизации в RL - найти оптимальную стратегию π^*, которая оптимизирует общее накопленное вознаграждение.

2.4 Оптимизация политики

Актор-критик преодолевает разрыв между методами аппроксимации функции стоимости и градиентными методами политики для RL. Этот подход доказал свою способность обучаться и адаптироваться к большим и сложным средам, и был использован для популярных видеоигр, таких как Doom [Wu and Tian(2017)]. Таким образом, акторно-критический подход применяется для торговли с большим портфелем акций [Xiong et al.(2018)]. Алгоритм Proximal Policy Optimization (PPO) (Schulman et al., 2017 [Schulman et al.(2017)]) представляет собой акторно-критический внеполитический безмодельный RL-алгоритм для дискретного управления и непрерывного управления. Однако он сталкивается с серьезной неэффективностью выборки и требует огромного объема выборки для обучения, что неприемлемо для обучения на реальном рынке фондовой торговли. Другим типичным примером алгоритмов с акторной критикой является алгоритм Deep deterministic policy gradient (DDPG) (Lillicrap et al., 2015 [Lillicrap et al.(2015)]). Исследователи изучают потенциал RL для биржевой торговли с помощью алгоритма DDPG и добиваются неплохих результатов [Emami(2016), Azhikodan et al.(2019)]. Хотя он очень чувствителен к различным гиперпараметрам во время обучения, поэтому отличные результаты DDPG на различных бенчмарках на самом деле являются искусственными, и его сложно использовать для решения многих конкретных задач.

Туомас Хаарноя предложил алгоритм SAC [Haarnoja et al.(2018)], метод вне политики, для решения некоторых проблем неэффективности выборки в PPO и чувствительности гиперпараметров в DDPG. Самым большим отличием SAC является его цель максимизации энтропии, что является большим преимуществом для поддержания стохастичности при выборе политики, и, следовательно, может помочь агенту искать пространство состояний более полно, чтобы избежать локального оптимума и чувствительности.

Для случайной величины x с плотностью вероятности p энтропия $H\,x$ может быть определена как:

$$H(p) = \mathop{\mathrm{E}}_{x \sim p}[-\log p(x)] \tag{6}$$

Если рассматривать энтропийную регуляризацию RL, то к вознаграждению будет добавлен дополнительный член, который является энтропией. В то же время влияние текущего действия на будущее вознаграждение со временем ослабевает, поэтому вместо направленного вознаграждения мы

используем дисконтированное вознаграждение. Мягкая функция ценности (мягкая функция Q) становится:

$$Q^{\pi}_{\text{мягкий}}(s, a) = \underset{(s_t, a_t)\sim\rho_\pi}{\mathrm{E}} [\Sigma_{t=0}^{\infty} \gamma^t \, r(s_t, a_t) + \alpha \cdot \Sigma_{t=1}^{\infty} \gamma^t \, H(\pi(\cdot|s_t)) | s_0 = s, a_0 = a] \quad (7)$$

где ρ_π представляет собой распределение пары "состояние-действие" агента в рамках политики π. Соответственно, мягкая функция V становится:

$$V^{\pi}_{\text{мягкий}}(s) = \underset{(s_t, a_t)\sim\rho_\pi}{\mathrm{E}} [\Sigma_{t=0}^{\infty} \gamma^t \, (r(s_t, a_t) + \alpha H(\pi(\cdot|s_t))) | s_0 = s] \quad (8)$$

Согласно этим определениям, V^{π} и Q^{π} связаны между собой

$$V^{\pi}(s) = \mathrm{E}_{a\sim\pi}[Q^{\pi}(s, a) + \alpha H(\pi(-/s))]. \quad (9)$$

Таким образом, уравнение Беллмана можно записать в виде:

$$Q^{\pi}_{\substack{\text{мяг}\\ \text{кий}}}(s, a) = \mathrm{E}_{s'\sim P(s'|s,a)}[r(s, a) + \gamma(Q^{\pi}_{\substack{\text{мяг}\\ \text{кий}}}(s', a') + \alpha H(\pi(-/|s')))] \quad (10)$$

За последние г о д ы существует несколько версий SAC. Здесь мы используем более старую версию, в которой в дополнение к функции Q изучается функция стоимости V.

$$Q^{\pi}_{\substack{\text{мяг}\\ \text{кий}}}(s, a) = \mathrm{E}_{s'\sim P(s'|s,a)}[r(s, a) + \gamma(V^{\pi}_{\substack{\text{мяг}\\ \text{кий}}}(s')] \quad (11)$$

Таким образом, итерация функции стоимости

$$V^{\pi}_{\text{мягкий}}(s) = \mathrm{E}_{a\sim\pi}[Q^{\pi}_{\text{мягкий}}(s, a) - \alpha log\pi(a/s)]. \quad (12)$$

С помощью трюка с обрезанным донатом Q функция потерь Q-сети в SAC имеет вид:

$$L(\varphi, D) = \mathrm{E}_{(s,a,r,s')\in D)}[Q_{\varphi}(s, a) - y(r, s', d)]^2 \quad (13)$$

где Q_{φ} и $_{Qtarget}$ - нейронные сети и

$$y(r, s', d) = r + \gamma(1 - d)(\min_{j=1,2} Qtarget_j(s', \tilde{a}') - \alpha log\pi_{\theta}(\tilde{a}'/s')) \quad (14)$$

где $\tilde{a}' \sim \pi_{\theta}(-/s')$, в котором выборка из $\pi_{\theta}(-/s)$ берется по гауссовскому распределению.

Теперь мы сформулируем эту задачу как стохастическую задачу оптимального управления, а стохастическая задача оптимального управления может быть записана в виде:

$$J^{u}(x) = \mathrm{E}[\int_0^{\infty} f(X_s, u_s)e^{-\gamma s} ds/X^{u} = x] \quad (15)$$

ограниченный уравнением (5).
Определим

$$V(x) = \inf_{u} J^{u}(x) - \alpha log\pi(a/s) \quad (16)$$

Тогда V удовлетворяет зависящему от времени уравнению HJB:

$$\inf_{u} L^{u} V(x, u) + f(x, u) - \gamma V(x) = 0 \quad (17)$$

где $L^{u} V(x) = \frac{1}{2} Tr(\sigma\sigma^{T} Hess(V))(x, u) + b(x, u)^{T} \nabla V(x)$ - генератор из SDE (5).

3 Брак между MBRL и RSRS

В этой статье мы используем индикатор рыночного тайминга Resistance Support Relative Strength (RSRS) [1]. Это технический индикатор, позволяющий выбрать лучшее время для покупки и продажи путем измерения относительной силы поддержки и сопротивления [Lloyd Sr(2013)]. Используя определения уровней поддержки и сопротивления в финансовой сфере, мы имеем:

• **Уровень сопротивления**: Цена, при которой, по мнению трейдера, сила продавца начинает преобладать над силой покупателя, что затрудняет дальнейшее движение цены.
подниматься или отступать от падающей цены.

• **Уровень поддержки**: Цена, при которой, по мнению трейдеров, сила покупателя начинает преобладать над силой продавца, тем самым останавливая падение или отскок
растущая цена.

Индикатор RSRS больше не рассматривает сопротивление и поддержку как фиксированное значение, а как переменную в этом отчете. Этот индикатор представляет собой ожидаемое суждение трейдеров о верхней и нижней границах текущего состояния рынка.

Исследователи заменили конкретный ценовой порог поддержки и сопротивления на относительную силу изменения дневных максимумов и минимумов цены. Пусть μ - среднее значение исторического наклона RSRS, а σ - стандартное отклонение
исторический наклон. Обозначим уровень сопротивления как S_{buy} , где $S_{buy} = \mu + \sigma$. Затем обозначьте уровень поддержки как S_{sell} , где $S_{sell} = \mu - \sigma$. Если наклон RSRS больше S_{buy} , покупаем всю позицию, а если меньше S_{sell} , продаем.
и закройте позицию. Таким образом, можно контролировать просадку и гарантировать прибыльность стратегии.

3.1 Индикатор RSRS

Индикатор RSRS использует величину изменения высокой цены при изменении низкой цены определенной акции на 1 единицу для измерения силы поддержки и сопротивления. То есть регрессия ряда высокой цены и ряда низкой цены за определенный период. Наклон β, полученный с помощью модели, является специфическим показателем для измерения относительной силы поддержки сопротивления.

Берем временной ряд высокой цены и временной ряд низкой цены за предыдущие N дней по определенной акции. Вычисляем наклон RSRS за день:

$$высокий = \alpha + \beta \times низкий + \epsilon \tag{18}$$

где $\epsilon \sim N(0, \sigma)^2_\epsilon$

3.2 Правильный стандартный балл

Для акций, находящихся в разных периодах рынка, среднее значение наклона будет сильно колебаться. Поэтому не стоит напрямую использовать среднее значение наклона в качестве индекса времени. Хорошим выбором будет стандартизация значения наклона.

[1] Junwei Liu, Xiaoxiao Zhou. Market Timing Based on Resistance Support Relative Strength (RSRS). *Everbright Securities Technical Timing Series Report Series I*, 2017.5.1

1) Берем временной ряд наклона RSRS за предыдущие M дней по определенной акции. Вычисляем стандартный показатель $RSRS_{std}$ наклона RSRS за день:

$$RSRS_{std} = \frac{RSRS - \mu_M}{\sigma_M} \tag{19}$$

где μ_M - средний наклон за предыдущие M дней, а σ_M - стандартное отклонение за предыдущие M дней.

2) На самом деле, когда наклон используется для количественной оценки относительной силы поддержки сопротивления, его количественный эффект в значительной степени зависит от эффекта подгонки. Поэтому далее мы учитываем эффект подгонки и взвешиваем стандартные оценки с помощью коэффициента детерминации (значение R-квадрат в регрессионной модели), чтобы уменьшить влияние стандартной оценки RSRS, которая имеет большое абсолютное значение, но плохой эффект подгонки, на стратегию.

$$RSRS_{cor} = RSRS_{std} \times R^2 \tag{20}$$

где R^2 - коэффициент детерминации в регрессионной модели дня.

3) Индикатор RSRS сам по себе обладает отличной способностью к левостороннему прогнозированию. Поэтому в данной работе мы используем значительную поправку на правостороннюю стандартную оценку в качестве временного индекса.

$$_{RSRSrightdev} = RSRS_{cor} \times RSRS \tag{21}$$

4) Согласно проведенному ранее анализу, временной индекс RSRS определяется как:

$$RSRS_{rightdev} = \frac{RSRS - _{\mu M}}{\sigma_M} \times R^2 \times RSRS \tag{22}$$

3.3 Предлагаемая модель

Основываясь на индикаторе RSRS, описанном в предыдущей сессии, мы знаем, что это относительно сильная политика, позволяющая противостоять глубокому падению, когда акции сильно падают в течение какого-то срочного периода времени. Поэтому мы считаем, что сочетание этой стратегии с классическими алгоритмами RL может помочь получить оптимальную стратегию с меньшей волатильностью и большей стабильностью, что означает, что мы можем получить более высокий коэффициент Шарпа и меньшую максимальную просадку. Среди всех современных RL-алгоритмов, основанных на моделях, мы выбрали MBPO и M2AC в качестве наших иллюстрированных базовых моделей, хотя мы считаем, что другие модели, такие как BMPO [Lai et al.(2020)] и MVE [Feinberg et al.(2018)], также могут иметь похожие результаты. Псевдокод приведен в таблице 2, где в качестве примера используется MBPO, а случай с M2AC аналогичен и прост в получении. Для наглядности далее мы обозначим алгоритм MBPO с индикатором RSRS как стратегию RSPO, а алгоритм M2AC с индикатором RSRS как

стратегию RSAC.

4 Эксперимент

В этой сессии мы представим некоторые результаты экспериментов и соответствующие экс-планации, из которых можно увидеть, как и почему RSRS помогает улучшить

Таблица 2: Рамочная программа RSPO

Алгоритм RSPO: алгоритм MBPO с индикатором RSRS

Инициализируйте политику π_θ , динамическую модель P_φ , набор данных окружения D_{env} , набор данных модели D_{model}

Инициализируем временное окно l, рассматриваем окно M, порог покупки rs_{buy} , порог продажи rs_{sell} и максимальный акт

для E эпох сделайте

Обучите динамическую модель P_φ на D_{env}

для N шагов

$\rightarrow y = \alpha + \beta_i - \rightarrow x + \epsilon$, ($i = 1, ..., N$), где $\rightarrow y$ и $\rightarrow x$ соответствуют последовательности высоких и низких цен длины l соответственно.

для M шагов сделайте

$\rightarrow y = \alpha + \beta_m - \rightarrow x + \epsilon$, ($m = i - M + 1, ..., i$)

$\beta stdi = \frac{\beta i - E(\beta_j)}{\sigma(\beta_j)}$, $j = i - M + 1, ..., i$

$\beta_{mod}\, i = \beta_{std}\, i \times R^2$, где R^2 - коэффициент детерминации, соответствующий

$\beta_{i\, \beta rightdevi} = \beta mod \times \beta i$

Получите действие a_t , используя политику π_θ

если $_{\beta rightdevi} > rs_{buy}$ **, то** $a_t = hmax$

если $_{\beta rightdevi} < rs_{sell}$ **, то** $a_t = - hmax$

Выполните действие a_t в окружении ;

добавьте в D_{env} **для L модельных разворотов do**

Выборка s_t случайным образом из D_{env}

Выполните k-шаговое развертывание модели, начиная с s_t , используя политику π_θ ; добавьте в D_{model}

для W градиентных обновлений сделайте

Обновление параметров политики на D_{model} : $\vartheta \leftarrow \vartheta - \lambda_\pi \nabla_\theta (J_\pi (\vartheta, D_{model}))$

эффективность классических моделей обучения с подкреплением в портфеле акций.

4.1 Набор данных

Все данные для выборки взяты из базы данных Yahoo finance [2]. Мы проводим эксперименты на 30 акциях, отобранных по уровню их оборачиваемости в течение 180 дней до 01 января 2009 года на рынке Standard and Poor's 500 (S&P 500). Отобранные акции имеют относительно низкую скорость оборота, чтобы избежать большого шума или колебаний

на нестабильных финансовых рынках с небольшой рыночной капитализацией. В данном случае, в качестве показателя ликвидности, "скорость оборота" означает частоту перехода акций из рук в руки в

[2] https://www.yahoo.com

рынок в течение определенного периода времени. Таким образом, мы сможем лучше изучить применимость и универсальность предложенного в данной работе агента.

Для каждого эксперимента в качестве временного диапазона набора данных используются исторические ежедневные ценовые данные с 1 января 2009 года по 3 июля 2021 года. Данные с 1 января 2009 года по 3 июля 2016 года (**1888 дней**) используются в качестве обучающего набора, а данные с 4 июля 2016 года по 3 июля 2018 года (**504 дня**) - в качестве проверочного набора. Оставшиеся данные с 4 июля 2018 года по 3 июля 2021 года (**755 дней**) используются в качестве тестового набора. Мы обучаем нашего агента на обучающих данных, затем выбираем гиперпараметры модели, оценивая такие показатели, как годовая доходность, коэффициент Шарпа и максимальная просадка на валидационных данных, и, наконец, применяем выбранную модель на тестовых данных, что представлено на следующих рисунках и таблицах. Обратите внимание, что все результаты тестирования усреднены по 10 случайным экспериментам.

Здесь мы задаем следующие гиперпараметры для наших экспериментов: начальный баланс B_0 = 1e6 долларов, временное окно l = 10, временной период среднего и стандартного отклонения для RSRS M = 300, два порога rs_{buy} = 1.0, rs_{sell} = -0.4, и
максимальное количество акций в одной сделке $hmax$ = 100.

4.2 Показатели портфеля

В наших экспериментах используются пять алгоритмов, PETS, MBPO, M2AC, RSPO, RSAC. Все эти стратегии на тестовых данных сравниваются с базовым индексом рынка: GSPC.

Для оценки эффективности предложенной модели мы используем следующие показатели.

- Годовая доходность: среднее геометрическое количество денег, зарабатываемых инвестиционной стратегией каждый год в течение определенного периода времени.
- Кумулятивная доходность: отражает общий эффект от торговой стратегии в определенном временном диапазоне.
- Годовая волатильность: годовое стандартное отклонение доходности портфеля показывает устойчивость агента.
- Коэффициент Шарпа [Sharpe(1998)]: доход, полученный на единицу волатильности, который является широко используемым показателем эффективности инвестиций.
- Коэффициент Кальмара: годовая доходность, полученная на единицу максимальной просадки за период.
- Стабильность: Определяет R-квадрат линейной подгонки кумулятивного логарифма доходности.

- Максимальная просадка: максимальный убыток от снижения инвестиций от пика к корыту до достижения нового пика.

Общая производительность при различных измерениях: В таблице 3 мы используем семь вышеуказанных показателей для оценки эффективности различных алгоритмов RL. Из таблицы видно, что в период с июля 2018 года по июль 2021 года все алгоритмы обучения с подкреплением на основе моделей превосходят базовый метод

Таблица 3: Эффективность стратегии в S&P 500. Конкретные показатели шести стратегий по семи бэктестовым индикаторам за трехлетний период бэктестирования с июля 2018 года по июль 2021 года.

	RSAC	RSPO	M2AC	MBPO	ДОМАШНИЕ ЖИВОТНЫЕ	GSPC
В годовом исчислении Возврат	**30.02%**	28.23%	23.43%	22.78%	20.21%	16.75%
Кумулятивный Возврат	**119.56%**	110.62%	87.87%	84.96%	73.60%	59.04%
В годовом исчислении Волатильность	**22.78%**	23.05%	23.58%	24.91%	24.15%	23.05%
Шарп Соотношение	**1.27**	1.2	1.01	0.95	0.88	0.79
Калмар Соотношение	105.14%	**106.86%**	72.57%	64.58%	61.11%	49.37%
Стабильность	94.63%	**95.54%**	89.08%	88.96%	90.03%	72.82%
Максимальный Просадка	-28.55%	**-26.41%**	-32.28%	-35.28%	-33.07%	-33.92%

(GSPC), особенно RSAC и RSPO, которые занимают 2 первых места по всем показателям. Если говорить более конкретно, то, во-первых, с точки зрения прибыли, RSAC имеет годовую доходность около 30 % и кумулятивную доходность 120 %, что почти в два раза выше, чем у GSPC. Во-вторых, с точки зрения управления рисками, RSAC и RSPO имеют относительно низкую годовую волатильность, которая сочетается с высоким коэффициентом стабильности и низкой максимальной просадкой, что означает, что RSRS может помочь агенту принять более стабильное инвестиционное решение с меньшей степенью риска. И, наконец, если рассматривать управление рисками и получение прибыли, коэффициент Шарпа и коэффициент Калмара показывают, что RSAC и RSPO значительно лучше других, что указывает на то, что сочетание RSRS с некоторыми классическими алгоритмами обучения с подкреплением является достойным направлением для опробования.

Сравнение годовой доходности: В таблице 4 показана годовая доходность шести стратегий за каждый год. Анализ табличных данных показывает, что в период с июля 2019 года по июль 2020 года из-за влияния COVID-19 настроения на фондовом рынке были очень вялыми. Доходность эталонного GSPC составляет всего 4,66%, а M2AC, MBPO, PETS, как наши

контрольные группы, имеют максимальную доходность менее 10%. Однако предложенный нами метод с использованием RSRS в M2AC и MBPO все еще сохраняет годовую доходность около 20% и 28%, показывая, что индикатор RSRS может эффективно противостоять падению, когда финансовый рынок терпит крах во время непредсказуемого кризиса.

Однако с июля 2020 года по июль 2021 года, когда последовало быстрое восстановление фондового рынка, доходность M2AC и MBPO резко выросла, опередив RSAC и RSPO. Мы предполагаем, что причиной этого может быть низкий объем активов M2AC и MBPO в июле 2020 года.

Исходя из приведенного выше анализа, мы считаем, что алгоритмы M2AC и MBPO сравнительно чувствительны к настроениям рынка, что свидетельствует о том, что RSRS

Таблица 4: Годовая доходность шести стратегий в каждом году с июля 2018 по июль 2021.

	2018.7.1-2019.7.1	2019.7.1-2020.7.1	2020.7.1-2021.7.1
RSAC	**29.90%**	**19.53%**	**38.33%**
RSPO	**31.72%**	**27.88%**	22.66%
M2AC	23.22%	5.94%	**42.25%**
MBPO	22.46%	8.01%	36.94%
ДОМА	15.42%	9.31%	34.13%
ШНИЕ			
ЖИВО			
ТНЫЕ			
GSPC	9.51%	4.66%	36.87%

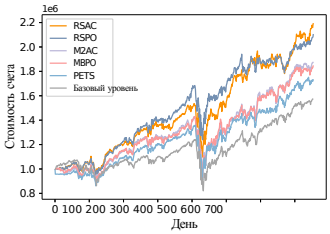

Рисунок 1: Доходность. Кривая "время-доходность" RSAC, RSPO и M2AC, MBPO, PETS, базовый уровень.

индикатор может в определенной степени повысить устойчивость алгоритмов к резким колебаниям финансового рынка.

Сравнение ежедневной доходности: Для всех 30 отобранных акций мы показываем портфель пяти алгоритмов и соответствующую базовую линию на рисунке 1. Он и л л ю с т р и р у е т значение счета шести стратегий в эксперименте. Сравнивая показатели RSAC, RSPO, M2AC и MBPO, можно заметить, что после добавления индикатора RSRS доход RSAC (оранжевый) и RSPO (темно-синий) значительно увеличился и стабильно превышает кумулятивный доход PETS (светло-голубой) и базовой линии (серый).

4.3 Производительность модели

Сходимость сети критиков: Чтобы гарантировать надежность наших алгоритмов, мы также анализируем сходимость критической сети. На рисунке 2 показана зависимость между потерями критика от количества обучающих эпизодов, из которой следует, что все алгоритмы сходятся менее чем за 50 эпизодов. Кроме того, картина снижения потерь критики в течение первых 30 обучающих эпизодов показывает, что RSAC (оранжевый) и RSPO (темно-синий) сходятся быстрее, чем чистые MBPO (фиолетовый) и M2AC (розовый) на ранней стадии, что указывает на то, что RSRS помог нашему предложенному методу быть

относительно эффективным с точки зрения вычислительных затрат. Меньшая рамка на рисунке 2 показывает сходимость в деталях. Кроме того, на рисунке 3 показан логарифм потерь критика в тренировочных эпизодах, и мы видим, что потери критика RSAC (оранжевый) и RSPO (темно-синий) имеют стабильно более низкую ошибку по сравнению с

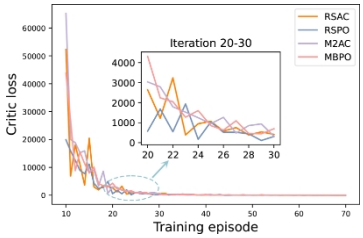

Рисунок 2: Критические потери при обучении .

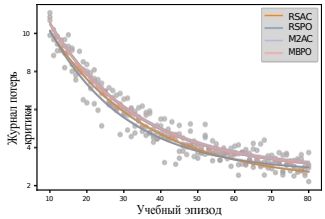

Рисунок 3: Критические потери при обучении .

с MBPO и RSPO без RSRS.

Анализ динамической модели перехода: Все алгоритмы RL, используемые в этой статье, основаны на моделях, поэтому мы также проанализируем производительность наших динамических моделей. При обучении переходной динамики с использованием четырехкратных данных

(s, a, s', r), полученных из реальной среды, для любого заданного (s, a) мы можем предсказать s' в виртуальной среде, и поэтому мы также проверяем сходимость переменной s' . Поскольку наше состояние имеет высокую размерность, в качестве иллюстрации мы возьмем некоторые компоненты состояния 's'. Например, на рисунке 4 показана абсолютная ошибка между предсказанное "следующее состояние" и истинное "следующее состояние", спроецированное на координату "кумулятивные доли" (обозначены как W_t в таблице 1). Видно, что с увеличением числа обучающих итераций предсказанные доли из динамики переходов приближаются к истинным долям, заданным реальной средой. Кроме того, на рисунке 5 показана ошибка, спроецированная на компонент "баланс", который представляет собой денежные средства на счете с течением времени (обозначен как B_t в таблице 1).

5 Заключение

В этой статье мы предлагаем новую стратегию, которая объединяет индикатор поддержки сопротивления и относительной силы (RSRS) с некоторыми мощными методами RL, основанными на модели, для

оптимизации торговли акциями.

Стратегии RSAC и RSPO используют преимущества как временной селекции из технического анализа на финансовом рынке, так и свойства интеллектуального взаимодействия агента и среды, присущие алгоритмам RL на основе моделей, что приводит...

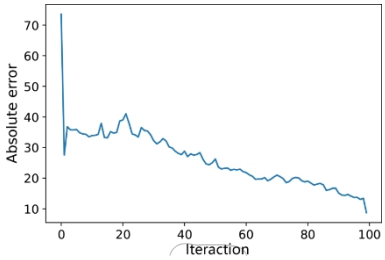

Рисунок 4: Абсолютная ошибка кумулятивных долей

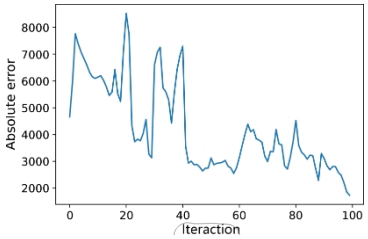

Рисунок 5: Абсолютная погрешность баланса

В результате улучшаются показатели стабильности, управления рисками и получения прибыли, особенно устойчивость к крупным маркетинговым кризисам. Результаты экспериментов на 30 акциях, отобранных из рынка S&P 500, показывают, что предложенные стратегии RSAC и RSPO могут эффективно усилить производительность алгоритмов, основанных на модели. Например, в период пандемии COVID-19 наши стратегии значительно улучшают показатели максимальной просадки и годовой доходности. Чтобы гарантировать достоверность наблюдаемых результатов, мы также проверяем сходимость критических потерь агента и ошибки предсказания динамической модели. Таким образом, мы убедились, что объединение методов из области финансов и статистики с алгоритмами RL является интересным направлением исследований.

Тем не менее, перед нами все еще стоят некоторые задачи. Например, как известно, существует тесная связь между RL и стохастическим оптимальным управлением. После постановки этого вопроса как задачи стохастического оптимального управления, как доказать единственность решения и его свойства в соответствующем функциональном пространстве, нам пока неизвестно. Более того, анализ границы ошибки обобщения РЛ на основе моделей также важен для разработки более адаптивных и устойчивых алгоритмов РЛ.

Ссылки

[Appel(2003)] Gerald Appel. 2003. Станьте своим собственным

техническим аналитиком: Как определить важные поворотные точки рынка с помощью скользящей средней

индикатор конвергенции-дивергенции или macd. *The Journal of Wealth Manage- ment* 6, 1 (2003), 27-36.

[Azhikodan et al.(2019)] Akhil Raj Azhikodan, Anvitha GK Bhat, and Ma- matha V Jadhav. 2019. Бот для торговли акциями с использованием глубокого обучения с подкреплением. In *Innovations in Computer Science and Engineering*. Springer, 41-49.

[Bao et al.(2017)] W. Bao, J. Yue, Y. Rao, and P. Boris. 2017. Система глубокого обучения для финансовых временных рядов с использованием стековых автоэнкодеров и долгосрочной и краткосрочной памяти. *PLoS ONE* 12, 7 (2017), e0180944.

[Bollinger(1992)] Джон Боллинджер. 1992. Использование полос Боллинджера. *Stocks & Com- modities* 10, 2 (1992), 47-51.

[Цай и Вэй (2020)] Т. Тони Цай и Хунчжи Вэй. 2020. Распределенная оценка среднего гаус-сиана при коммуникационных ограничениях: Оптимальные скорости и коммуникационно эффективные алгоритмы. *arXiv препринт arXiv:2001.08877* (2020).

[Chua et al.(2018)] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. 2018. Глубокое обучение с подкреплением за несколько попыток с использованием моделей вероятностной динамики. *arXiv preprint arXiv:1805.12114* (2018).

[Эмами (2016)] Патрик Эмами. 2016. Глубокие детерминированные градиенты политики в tensorflow. *Мои резюме работ по машинному обучению и исследованиям различных тем, касающихся искусственного интеллекта* (2016).

[Feinberg et al.(2018)] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. 2018. Расширение значений на основе модели для эффективного безмодельного обучения с подкреплением. В *материалах 35-й Международной конференции по машинному обучению (ICML 2018)*.

[Haarnoja et al.(2018)] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Мягкий актор-критик: Внеполитическое максимально энтропийное глубокое обучение с подкреплением и стохастическим актором. *Международная конференция по машинному обучению*. PMLR, 1861-1870.

[Janner et al.(2019)] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. 2019. Когда доверять своей модели: Оптимизация политики на основе модели. *arXiv preprint arXiv:1906.08253* (2019).

[Kang et al.(2018)] Qinma Kang, Huizhuo Zhou, and Yunfan Kang. 2018. Асинхронный метод акторно-критического обучения с подкреплением для выбора акций и управления портфелем. In *the 2nd International Conference*.

[Lai et al.(2020)] Hang Lai, Jian Shen, Weinan Zhang, and Yong Yu. 2020. Двунаправленная оптимизация политики на основе моделей. *Международная конференция по машинному обучению*. PMLR, 5618-5627.

[Lee et al.(2019)] J. Lee, R. Kim, Y. Koh, and J. Kang. 2019. Global Stock Market Prediction Based on Stock Chart Images Using Deep Q-Network. *IEEE Access* PP, 99 (2019), 1-1.

[Li et al.(2019a)] X Li, Y. Li, Y. Zhan, and X. Y. Liu. 2019a. Optimistic Bull or Pessimistic Bear: Adaptive Deep Reinforcement Learning for Stock Portfolio Allocation. *Papers* (2019).

[Li et al.(2019b)] Y. Li, M. Nee, and V. Chang. 2019b. Эмпирическое исследование инвестиционной стратегии фондового рынка на основе модели глубокого обучения с подкреплением. In *4th International Conference on Complexity, Future Information Systems and Risk (COMPLEXIS 2019)*.

[Lillicrap et al.(2015)] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Непрерывное управление с глубоким обучением с подкреплением. *arXiv preprint arXiv:1509.02971* (2015).

[Lloyd Sr(2013)] Tom K Lloyd Sr. 2013. *Successful Stock Signals for Traders and Portfolio Managers: Integrating Technical Analysis with Fundamentals to Improve Performance*. John Wiley & Sons.

[Mnih et al.(2016)] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Асинхронные методы глубокого обучения с подкреплением. *Международная конференция по машинному обучению*. PMLR, 1928-1937.

[Mnih et al.(2013)] Владимир Мних, Корай Кавуккуоглу, Дэвид Сильвер, Алекс Грейвс, Иоаннис Антоноглу, Даан Виерстра и Мартин Ридмиллер. 2013. Игра в Atari с глубоким обучением с подкреплением. *arXiv препринт arXiv:1312.5602* (2013).

[Mnih et al.(2015)] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Управление на уровне человека с помощью глубокого обучения с подкреплением. *nature* 518, 7540 (2015), 529-533.

[Pan et al.(2020)] Feiyang Pan, Jia He, Dandan Tu, and Qing He. 2020. Доверять модели, когда она уверена в себе: Критика актора на основе маскированной модели. *Advances in neural information processing systems* 33 (2020), 10537-10546.

[Pastore et al.(2016)] A. Pastore, U. Esposito, and E. Vasilaki. 2016. Модификация инвесторов фондового рынка как агентов обучения с подкреплением. *Международная конференция IEEE по эволюционирующим* и *адаптивным интеллектуальным системам*.

[Peng et al.(2018)] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam- Fai Wong, and Shang-Yu Su. 2018. Deep dyna-q: интеграция планирования для обучения политике диалога по выполнению задачи. *arXiv preprint arXiv:1801.06176* (2018).

[Schulman et al.(2017)] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Алгоритмы оптимизации проксимальной политики. *arXiv preprint arXiv:1707.06347* (2017).

[Sharpe(1998)] William F Sharpe. 1998. Коэффициент Шарпа. *Streetwise-the Best of the Journal of Portfolio Management* (1998), 169-185.

[Sornmayura(2019)] Сутта Сорнмаюра. 2019. Надежная торговля на рынке Форекс с помощью глубокой q-сети (dqn). *ABAC Journal* 39, 1 (2019).

[Ţăran-Moroşan(2011)] Адриан Ţăran-Moroşan. 2011. Пересмотр индекса относительной силы. *African Journal of Business Management* 5, 14 (2011), 5855-5862.

[Van Hasselt et al.(2016)] Хадо Ван Хассельт, Артур Гез и Дэвид Сильвер. 2016. Глубокое обучение с подкреплением и двойным q-обучением. В *материалах конференции AAAI по искусственному интеллекту*, том 30.

[Wang et al.(2016)] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. 2016. Дуэльные сетевые архитектуры для глубокого обучения с подкреплением. *Международная конференция по машинному обучению*. PMLR, 1995-2003.

[Wu and Tian(2017)] Yuxin Wu and Yuandong Tian. 2017. Обучающий агент для игры в шутер от первого лица с акторно-критическим обучением. In *ICLR*.

[Xiong et al.(2018)] Zhuoran Xiong, Xiao-Yang Liu, Shan Zhong, Hongyang Yang, and Anwar Walid. 2018. Практический подход к глубокому обучению с подкреплением для биржевой торговли. *arXiv preprint arXiv:1811.07522* (2018).

[Yu et al.(2019)] Pengqian Yu, Joon Sern Lee, Ilya Kulyatin, Zekun Shi, and Sakyasingha Dasgupta. 2019. Глубокое обучение с подкреплением на основе моделей для динамической оптимизации портфеля. *arXiv preprint arXiv:1901.08740* (2019).