多変量時系列データを用いた分散型強化学習による低リスク行動の学習

Modeling Low-risk Actions from Multivariate Time Series Data using Distributional Reinforcement Learning

佐藤 葉介 [1*]　張 建偉 [2]
Yosuke Sato[1]　Jianwei Zhang [2]

[1] 岩手大学大学院総合科学研究科
[1] Graduate School of Arts and Science，Iwate University
[2] 岩手大学理工学部
[2] Faculty of Science and Engineering，Iwate University

Abstract: In recent years, investment strategies in financial markets using deep learning have attracted a significant amount of research attention. The objective of these studies is to obtain investment behavior that is low risk and increases profit. Although Distributional Reinforcement Learning (DRL) expands the action-value function to a discrete distribution in reinforcement learning which can control risk, DRL has not yet been used to learn investment action. In this study, we construct a low-risk investment trading model using DRL. This model is back-tested on Nikkei 225 data and compared with Deep Q Network (DQN). We evaluate the performance in terms of final asset amount, standard deviation, and the Sharpe ratio. The experimental results show that the proposed method can learn low-risk actions with the increasing profit, outperforming the compared method DQN.

1　はじめに

近年，深層学習を用いた金融市場に関する研究が盛んに行われている [1, 2]．金融市場は景気や為替などの経済的要因や政局などの経済外的要因など複雑な要因が関わり変動するため，確実な将来の状態予測や取引戦略の組み立てが困難な金融市場における投資行動の学習に関する研究はこれまでに多数されてきた [3, 4, 5, 6]．特に近年は高い特徴表現力を持つ深層モデルを用いて取引エージェントを製作する研究がされている [7, 8, 9]．多くの研究では利潤を増加させつつ保有する資産価値が減少するリスクへの対処をするという 2 つの課題に対して様々な手法が提案されてきた [10, 11, 12, 13]．ほとんどの深層強化学習を用いた先行研究では Deep Q Network(DQN) が提案手法のベースや比較手法として用いられている．これらの研究で用いられる評価値は，利潤の増大を測るためにテスト期間に得られた資産の多さ，その標準偏差をどれだけ安定して利益を得られるかを調べるために利用し，これら 2 つの評価値を統合した，取ったリスクに対してリターンの大きさを示すシャープレシオ [14] が主に用いられている．

一方，分散型強化学習 [15] は深層強化学習における行動価値関数の各行動の評価値を値の分布に拡張した手法であり，ベンチマークにおいて DQN，Double DQN(DDQN)，Dueling Network より優れた結果を残している．分散型強化学習による行動価値関数では，ある行動で得られる報酬の期待値だけでなく定義した報酬の値の範囲で，各報酬が得られる期待値を離散分布で学習することができる．モデルの出力を人が観察しリスク操作をすることが可能な利点があり，行動価値関数の出力と手動で設定した歪度の要素積を計算することで，ある程度学習の対象ごとの性質に合わせた行動の選択をする改善手法も提案されている [16]．

本研究では先に述べた金融市場の不確実性から起こる資産価値低下のリスクに対して，筆者の知るところで金融分野で応用されていない分散型強化学習を適用することで，低リスクな行動を獲得することを目的とする．日経 225 を構成する銘柄に対してバックテストを行い，DQN と比較してシャープレシオや得られた利益などについて評価を実施した．実験によりテスト環境の最終的な資産額についての標準偏差が，分散型強

*連絡先：岩手大学大学院総合科学研究科
〒 020-0066 岩手県盛岡市上田 4 丁目 3 - 5
E-mail: g0319083@iwate-u.ac.jp

化学習の方が小さく現れたことから日経 225 については DQN より小さなリスクの投資行動を学習できた結果が得られた．

2 関連研究

ポートフォリオマネジメントは利益の最大化やリスクの最小化を目的とした金融資産の分配をする問題であり，強化学習による適切な投資行動の学習が試みられている．価値ベースの手法として Shin[10] らは DQN を用いて 8 種類の暗号通貨と USD のポートフォリオマネジメントを学習させ，最も資産価値の減少を抑えた行動を学習できているため低リスクだとしている．方策ベースの手法として Xiong[11] らは Deep Deterministic Policy Gradient (DDPG) を用いて株価，保有株数，残高の状態からそれぞれの株に対する売却，保持，購入の行動を学習している．約 10 年分の Dow Jones 30 stocks についてバックテストを行い比較手法より利益とシャープレシオについて優位な結果を残した．また，Ye[13] らも DDPG を用いており，ニュース記事と株価の推移を前処理したデータからポートフォリオの割り当てを学習している．ベンチマークにおいては比較手法より優位な利益を出し，シャープレシオも概ね優位な結果を残した．CNN ベースの DDPG を構築し 12 種類の暗号通貨のポートフォリオマネジメントを行った研究もある [17]．Jiang[18] らは暗号通貨のポートフォリオマネジメントを行う EIIE フレームワークを開発し UBAH などの比較手法より portfolio value やシャープレシオについて優位な結果を得ている．EIIE フレームワークは方策ベースの手法であり，資産の潜在的な成長性から直近の予測を行う IIE のアンサンブル学習である．金融市場に関する研究では投資行動の学習に限らずリスクを考慮した手法が提案されているが，関連研究や過去の研究では分散型強化学習を用いた手法は取り入れられていない．本研究は分散型強化学習を投資行動の学習に初めて応用する．

3 手法

3.1 問題設定

本研究では金融市場における投資行動をマルコフ決定過程 $M(\mathcal{S}, \mathcal{A}, R, P)$ とする．$s \in \mathcal{S}$ は状態空間，$a \in \mathcal{A}$ は行動空間を表す．投資行動の対象資産は東京証券取引所第一部に上場している企業の株式と無リスク資産とし，所持している株式とともに状態 s に含まれる．また，行動 a は保持している無リスク資産で株式を購入，保持している株式の売却，資産の売買を行わない保持の 3 つを取りうる．

$R : S \times A \times S \to \mathbf{R}$ は報酬関数を表している．本手法では初期状態あるいは株式の購入時から売却時までの資産額の増減を報酬値に利用している．売却したときのみ即時報酬を与えるとそれまでの過程が評価されないため，報酬が得られたときに初期状態あるいは株式の購入を行ったときから売却したときまでの各状態 s に対して遅延報酬を与える．このとき，過去の状態に遡るにつれて割引率を適用することで偏りが現れる可能性を排除する．さらに各報酬値に対して DQN と同様に reward clipping を適用する．

遷移関数 $P : S \times A \times S \to [0, 1]$ は状態 s のとき行動 a を取り状態 s' へ遷移する状態遷移確率を表す．方策 $\pi(a \mid s)$ は状態 s の時の行動 a をとる確率を表す関数である．行動価値関数 $Q^\pi(s, a)$ は状態 s のとき方策 π に従い行動 a を取ったときに得られる期待報酬値を定義する．

$$Q^\pi(s, a) = \mathbf{E}[R(s, a)] + \gamma \mathbf{E}_{P,s}[Q^\pi(s', a')]$$

最適方策 π^* を学習し最適行動価値関数 $\mathbf{E}[Q^*(s, a)]$ の戻り値を最大化するような行動を学習することが目的となる．最適方策とは任意の初期状態 $s \in \mathcal{S}$ から期待報酬を最大化することである．最適方策の学習にはいくつか手法が存在するが，Q 学習では以下の更新式により最適行動価値関数を学習する．

$$Q_\theta(s, a) \leftarrow \mathbf{E}[R(s, a)] + \gamma \mathbf{E}_P[\underset{a'}{max} Q_\theta(s', a')]$$

DQN では Q_θ を CNN で表現しており様々な派生や応用 [9] がなされている．

3.2 分散型強化学習

森村 [19] らは期待リターンの再帰式であるベルマン期待方程式のリターンを分布に拡張した分布ベルマン方程式を定義している．分布ベルマン方程式を解くことでリターン分布を推定できるが分布ベルマン方程式は汎関数の自由度を持つため一般に推定は困難であるため近似が必要となる．

Bellemare[15] はリターン分布を多項分布で近似した categorical DQN を提案している．リターン分布は直感的には複数個の bin と呼ばれる 1 つの報酬値が得られる期待値を表すものが連続している．ハイパーパラメータとして設定した数だけの bin 数でリターン分布が構成される．近似リターン分布の bin 数 $M \geq 2$ と，近似リターン分布の上限 Q_{max} と下限 Q_{min} をハイパーパラメータとして定め，bin 間隔 Δ_z を

$$\Delta_z := \frac{Q_{max} - Q_{min}}{M - 1}$$

のように定数として定め，各 bin に対するリターン代表値を $Z_m, m \in \{1, ..., |m|\}$

$$Z_m := Q_{min} + (m-1)\Delta_z$$

とする．状態 s と行動 a の入力に対するリターン分布を表現する M 次元ベクトル $(q_1(s,a),..., q_M(s,a))$ を出力する深層モデル $\mathcal{S} \times \mathcal{A} \rightarrow \mathbf{R}^M$ を用いて，推定リターン分布 $\hat{P}$ を

$$\hat{P}(C = z_m|s,a) := \frac{exp(q_m(s,a))}{\sum_{m'=1}^{M} exp(q_{m'}(s,a))},$$

$$\forall m \in \{1, ..., M\}$$

として求める．このとき categorical DQN の行動価値の推定値 $\hat{Q}$ は

$$\hat{Q}(s,a) \triangleq \sum_{m=1}^{M} z_m \hat{P}(C = z_m|s,a)$$

となる．$\hat{Q}$ は DQN と同様に近似分布ベルマン行動最適作用素 $\hat{D}$ [15] を適用して現在の推定リターン分布 $\hat{P}$ から目的分布 $\hat{P}_n^{target}$ を求める．experience replay により得た経験 n を用いて目的分布 $\hat{P}_n^{target}$ と現在の推定分布 $\hat{P}(\cdot\,|s_n, a_n)$ との差異が小さくなるように深層モデルの重みを更新する．

学習は DQN と同様に experience replay を取り入れる [20]．学習において 1 ステップ更新されるごとに，replay memory に現在の状態，次の状態，評価値を保存する．n ステップに一度，ハイパーパラメータとして指定したバッチサイズだけ replay memory からデータを取り出し Target-Network の重みを学習する．

4 実験

categorical DQN による分散型強化学習を用いて，バックテストにより金融市場における投資行動を学習する実験を行った．DQN と比較して最終的な資産額，標準偏差，シャープレシオについて評価した．

4.1 データセットと前処理の手法

データセットは東京証券取引所第一部に上場しており，日経 225 に含まれる 225 銘柄を利用した．期間は 2010 年 1 月 4 日から 2019 年 12 月 30 日までの 10 年間の日足データを利用し，10 年分のデータが存在しない銘柄については，データが存在する年から 2019 年 12 月 30 日まで利用する．そのうちより過去の 9 割のデータを学習に用いて，より新しいデータを評価に用いる．市場は過去の状態の影響を受けて将来の状態が決定していると考えられるため，評価において未来の情報を学習していないモデルを用いるようにする．

日足には始値 (open price)，高値 (high price)，低値 (low price)，終値 (close price) の 4 つの変数が含まれ，それぞれの変数について前処理を行う．本手法では前日からの値動き，すなわち差分を学習させる．さらに DQN と同様に複数ステップの情報をまとめて 1 つの状態とする．あるステップ t における n 日分の時系列データを用いた場合に replay memory に保存するデータは以下のようになる．

$$v_t^{diff} = v_t - v_{t-1}$$

$$\boldsymbol{v_t} = (f(\{v_t^{open\ diff}), f(v_t^{high\ diff}), f(v_t^{low\ diff}), f(v_t^{close\ diff}\}))$$

$$\boldsymbol{s_t} = (\boldsymbol{v_t}, \boldsymbol{v_{t-1}}, ..., \boldsymbol{v_{t-n+1}})$$

v_t^{diff} はステップ t における前ステップとの差分の情報を含むベクトルを表す．$\boldsymbol{s_t}$ はステップ t における環境から観測される状態とする．f は引数として得た値に対して正規化を行う関数とする．

次状態 $\boldsymbol{s_{t+1}}$ はステップ t+1 において同様に計算したものである．r_t は 3.1 節で述べたように報酬が得られてから与えられるため，それまで $\{\boldsymbol{s_t}, a_t, r_t, \boldsymbol{s_{t+1}}\}$ の組をスタックする．報酬が得られてからスタックしたデータに割引率を適用した報酬を与え，replay memory に保存される．

4.2 日経 225 データのバックテストによる投資実験

実験環境は Open AI Gym をもとに構築した．モデルや分散型強化学習 (DRL) に利用するパラメータは固定し，225 銘柄それぞれについて環境を構築し投資行動の学習を行った．初期状態として投資エージェントは無リスク資産である ¥1,000,000 を所有する．環境から観測される状態は日足であり，4.1 節のように前処理を行い replay memory に保存する．学習データを用いて 1 epoch だけ replay memory を構築してから学習を開始し，設定した頻度で experience replay による学習を行う．次に評価用の replay memory を用いて評価用の未知の期間について投資行動を行い，リセットした初期資産から増減した資産額を求める．評価期間が終了したときにエージェントが株式を保持している場合は無リスク資産と株式の時価額の和を最終的な資産額とする．モデルの初期状態や方策が ε -greedy 法でありランダムな要素を含むため 100 回実験を行い，資産額の平均，標準偏差，シャープレシオを計算する．シャープレシオ [14] は評価期間の間に増加した資産額の平均をその標準偏差で割った値とした．実験では無リスク資産の利子率は 0 とする．

表 1: 225 銘柄のテスト結果の平均値

	DQN	DRL(ours)
Asset Amount(YEN)	1,037,772	**1,038,152**
Standard Deviation(YEN)	106,287	**100,938**
Sharpe Ratio	0.229	**0.239**

表 2: 225 銘柄のテスト結果の標準偏差

	DQN	DRL(ours)
Asset Amount(YEN)	119,477	**114,539**
Standard Deviation(YEN)	**83,549**	89,181
Sharpe Ratio	**0.702**	0.733

DRL と DQN の共通パラメータとして，予備実験により，モデルは 5 層全結合とし隠れ層数 32，Q 学習の割引率を 0.9，replay frequency を 4，Adam-ε を 1.5×10^{-4}，1 状態に含める日数を 8 日とした．DRL のパラメータは，モデルの最終層の bin 数を 31，Vmax を 10,Vmin を-10 とした．

評価値である最終的な資産額，標準偏差，シャープレシオについて期待値としての 225 銘柄の平均値を表 1 に，標準偏差を表 2 に示す．

表 1 より DRL は DQN に比べて 225 銘柄の平均の資産額，標準偏差，シャープレシオについて優位な結果となった．表 2 より 225 銘柄の各評価値の標準偏差は資産額のみ提案手法の方が優位な結果となっており，標準偏差やシャープレシオのばらつきは大きいことを示している．225 銘柄のうち最終的な資産額について DRL が優位であった銘柄の割合は 41.3%，標準偏差は 63.1%，シャープレシオは 53.3%の銘柄について優位な結果となった．シャープレシオの平均値については提案法は DQN の 1.04 倍のシャープレシオとなっており，DQN に比べて比較的低リスクであり大きな資産を得られるといえる [14].

5 考察

本実験では日経 225 に採用されているそれぞれの銘柄について学習し投資実験の評価を行った．この章では結果について考察する．

5.1 実験結果のヒストグラム

225 銘柄の結果について，最終的な資産のヒストグラムを図 1 に示す．いずれも縦軸は度数を表し，横軸については図 1 は各銘柄について最終的に得られた資産額の平均値，図 2 は最終的に得られた資産額の標準偏差，図 3 はシャープレシオを表している．図 1 が示

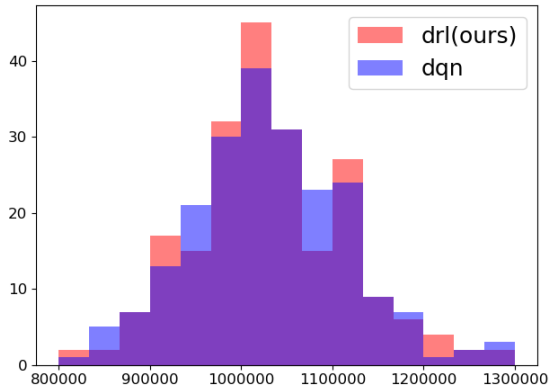

図 1: 最終資産額のヒストグラム比較

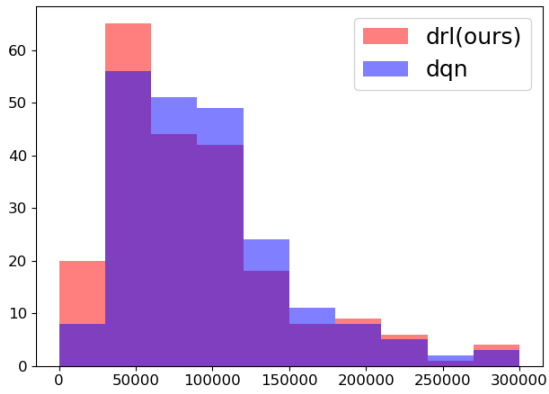

図 2: 最終資産額の標準偏差のヒストグラム比較

すように DRL と DQN が得られる最終的な資産額は目視だとほとんど差が無い．図 2 により，DQN に対して DRL は比較的多くの銘柄で標準偏差が小さいことから，比較的安定した取引行動を学習できているといえる．

5.2 有意性の検証

本研究では最終的な資産額，標準偏差，シャープレシオについて 225 銘柄それぞれの結果について平均と分散を計算しているが，結果の有意性について調べる．DRL により得られた結果と DQN により得られた結果を 2 つの群として代表値の有意性を検定する．

まず，DRL と DQN の結果についてシャピロ-ウィルク検定により標本に正規性があることを帰無仮説としテストした．DRL による結果の p 値は 5.49×10^{-10}，DQN による結果の p 値は 8.85×10^{-12} となり有意水準を 0.01 とすると帰無仮説が棄却され 2 つの結果は正規性を持たないことがわかった．

DRL と DQN の結果には同じ銘柄について投資行動を学習しているためデータ間の対応があり，正規性が

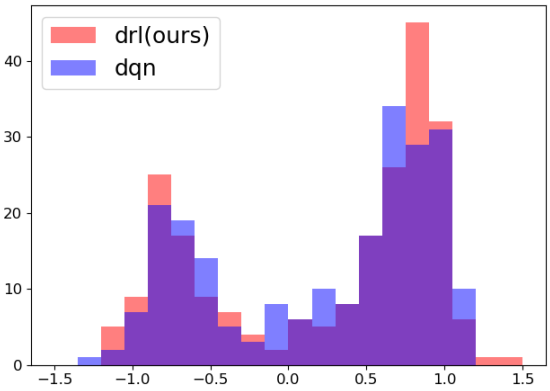

図 3: シャープレシオのヒストグラム比較

表 3: ウィルコクソンの符号付き順位検定結果

evaluation items	p value
Asset Amount	0.992
Standard Deviation	1.200×10^{-06}
Sharpe Ratio	0.470

無いことから，ノンパラメトリック検定であるウィルコクソンの符号付き順位検定により 2 群の標本の代表値に差がないことを帰無仮説としテストした．各評価項目に対する検定の結果を表 3 に示す．

有意水準を 0.01 とすると標準偏差については帰無仮説が棄却され 2 群の代表値には有意差があると言える．しかし最終的な資産額とシャープレシオについては統計的には有意差があるとは言えない結果となった．

5.3 epoch 数の妥当性

多くの実験では複数の epoch 数にわたって学習するが本実験では 1 epoch のみ学習している．予備実験としてランダムに抽出した銘柄コード 5301 について 10,000 epoch まで学習を行い，1 epoch ごとに評価を行った結果を図 4 に示す．損失関数の最適化手法は Adam を用いており，学習率は 6.25×10^{-5} と低めにしても，epoch 数の増加に従って評価における最終的な資産額は振動しており明確な増加はしなかった．

図 5 に 10,000 epoch まで学習したときのモデルが出力した Q 値の平均値と標準偏差の推移を表す．評価段階においてモデルが 1 step ごとに出力した Q 値を 10,000 epoch ほど累積したものについて，各 step で平均値と標準偏差を計算している．epoch 数が多くなるにつれて明らかな Q 値の増加は無く振動が続き，Q 値が蓄積されるにつれて標準偏差が大きくなっている．Q 値が増加し資産額が増加していれば報酬設計が間違っていることになるが，Q 値の明確な増加傾向が現れな

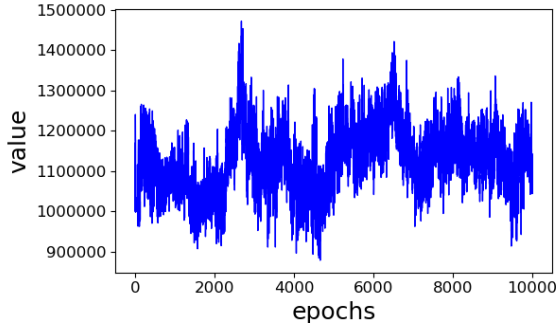

図 4: 分散型強化学習を用いた 1 epoch ごとの評価

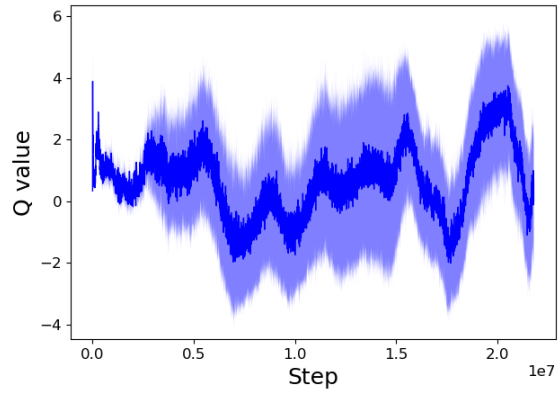

図 5: 分散型強化学習を用いた累計 Step ごとの Q 値

いことから epoch 数の増加に従い投資行動を改善するような学習がなされてないことがわかる．そのため 1 epoch のみの学習で評価を実施した．

6 結論

分散型強化学習により日経 225 を構成する銘柄からなる金融市場において投資行動の学習を行い，DQN と比較し最終的な資産額，標準偏差，シャープレシオの平均値について優位な結果を得た．標準偏差の結果は統計的な有意差があり，低リスクな行動を学習できている根拠となっている．また，資産額については銘柄ごとのばらつきが比較的小さく，DQN より安定した投資行動の学習が見込める．

今後の展望として，比較手法が少ないため多くの類似研究が採用している DDPG[21] などの手法との比較や LSTM に置き換えることが可能であると考えている．

参考文献

[1] Nicholas Tung Chan and Christian Shelton. An electronic market-maker. In *AI Memo 2001-005*,

2001.

[2] Y Hilpisch. Deep learning in finance. In *arXiv:1602.06561*, 2016.

[3] 石原龍太. 多層ニューラルネットワークと ga を用いた topix 運用 ai. 第 19 回人工知能学会 金融情報学研究会, 2017.

[4] 加藤旺樹, 穴田一. 遺伝的プログラミングを用いたテクニカル指標による金融取引の戦略木構築. 第 24 回人工知能学会 金融情報学研究会, 2020.

[5] 上田翼, 東出卓朗. 人工知能を用いた金融政策予想と市場予測分布に基づく為替の投資戦略. 第 18 回人工知能学会 金融情報学研究会, 2017.

[6] 宮坂純也, 穴田一. 心理的要素を考慮した投資行動モデル. 第 18 回人工知能学会 金融情報学研究会, 2017.

[7] Jia WU, Chen WANG, Lidong XIONG, and Hongyong SUN. Quantitative trading on stock market based ondeep reinforcement learning. In *IJCNN*, 2019.

[8] 小林弘幸, 和泉潔, 松島裕康, 坂地泰紀, 島田尚. 強化学習による高頻度取引戦略の構築. 第 24 回人工知能学会 金融情報学研究会, 2020.

[9] 常井祥太, 穴田一. Nt 倍率取引における深層強化学習を用いた投資戦略の構築. 第 22 回人工知能金融情報学研究会, 2019.

[10] Wonsup Shin, Seok-Jun Bu, and Sung-Bae Cho. Automatic financial trading agent for low-risk portfolio management using deep reinforcement learning. In *arXiv:1909.03278*, 2019.

[11] Zhuoran Xiong, Xiao-Yang Liu, Shan Zhong, Hongyang (Bruce) Yang, and Anwar Walid. Practical deep reinforcement learning approach for stock trading. In *NIPS 2018 Workshop on Challenges and Opportunities for AI in Financial Services*, 2018.

[12] Yifan Zhang, Peilin Zhao, Qingyao Wu, Bin Li, Junzhou Huang, and Mingkui Tan. Cost-sensitive portfolio selection via deep reinforcement learning. In *arXiv:2003.03051*, 2020.

[13] Yunan Ye, Hengzhi Pei, Boxin Wang, Pin-Yu Chen, Yada Zhu, and Bo Li Jun Xiao. Reinforcement-learning based portfolio management with augmented asset movement prediction states. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[14] W. Sharpe. The sharpe ratio. In *The Journal of Portfolio Management*, Vol. 1, pp. 49–58, 1994.

[15] Marc G. Bellemare, Will Dabney, and Remi Munos. A distributional perspective on reinforcement learning. In *ICML*, 2017.

[16] Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In *ICML*, 2018.

[17] Zhengyao Jiang and Jinjun Liang. Cryptocurrency portfolio management with deep reinforcement learning. In *Intelligent Systems Conference (IntelliSys) 2017*, 2017.

[18] Zhengyao Jiang, Dixing Xu, and Jinjun Liang. A deep reinforcement learning framework for the financial portfolio management problem. In *arXiv:1706.10059*, 2017.

[19] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation forreinforcement learning. In *InProceedings of the Conferenceon Uncertainty in Artificial Intelligence*, 2010.

[20] Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, M. G.Bellemare J. Veness, M. Riedmiller A. Graves, A. K. Fidjeland, G. Ostro-vski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, , and D. Hassabis. Human-level control through deep reinforcement learning. In *Nature*, pp. 529–533, 2015.

[21] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *arXiv:1509.02971*, 2015.