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

Trading System using Deep Reinforcement Learning

常井 祥太　　穴田 一

Shota Tokoi　　Hajime Anada

東京都市大学大学院工学研究科

Graduate School of Engineering, Tokyo City University

Abstract: In recent years, investment strategies using artificial intelligence have attracted a significant amount of research attention. However, it is difficult to construct an efficient investment strategy using artificial intelligence owing to the variable factors in market prices. Therefore, this study aims to focus on a trading method called the NT ratio transaction to reduce the number of price-variable factors. This transaction is an arbitrage transaction, which utilizes the difference in the price movements between Nikkei 225 futures and TOPIX futures. These futures generally exhibit similar price movements and even if the price differences expand, they tend to return to their original separation. Using this transaction, we can target profits from this price difference while offsetting a considerable number of price-variable factors. Therefore, in this study, we construct a model to acquire an investment strategy based on NT ratio transactions via deep reinforcement learning and confirm the effectiveness of this model.

1. はじめに

近年，人工知能に関する研究が活発に行われており, 例として, 強化学習[１]などを取り入れた将棋 AI である AlphaGo が人間のトップ棋士を破る大きな成果を残している．そのような中で，金融分野でも人工知能を用いた投資戦略の研究が行われている．松井らは複利型強化学習という新たな強化学習の枠組みを提案した．複利型強化学習とは，試行錯誤を通じてエージェントが将来獲得する報酬ではなく，複利式のリターン（得た利益を掛け金に上乗せして得るリターン）を最大化する行動規則を学習する枠組みである[２]. また, 彼らは複利型強化学習における行動価値関数をニューラル・ネットワークで表した複利型深層強化学習を提案した. この手法で彼らは, 日本国債の週次取引における行動規則を学習し，利益率が向上していく様子を確認した[３]. しかし, 最終的な利益率を見ると，学習が十分であるとは言い難い．これは国債や株価などには価格変動要因がかなり多く存在し，それらを十分に考慮できていないことが原因であると考えられる．しかし，これらは各国のニュースによる変動への影響など定量化が困難なものが多い．そこで，本研究では価格変動要因を減らすため, NT 倍率取引という取引手法に着目する．NT 倍率取引とは，日経 225 先物と TOPIX 先物の値動きの違いを利用した取引である．これらのような相関性の強い 2 つの金融商品に対して「買い」と「売り」をそれぞれ同時に行うことにより，価格の変動要因の大部分が相殺できるため，2 つの価格差のみに着目した取引が可能になる．また，松井らの手法では状態変数が 2 つと少なく，多数の状態変数を扱える深層強化学習の利点を活かし切れていない．そのため，状態変数を増やすことで，現在の状況を適切に捉えた上で，より良い投資行動を行えるようになるのではないかと考えた．さらに，通常の強化学習では，行動する度に報酬量を決定し付与するが，金融取引において，買う，売るなどの行動の良し悪しをすぐさま決定することは大変困難である. そこで，ポジションを取得してから保持し，解消するまでの行動に関する報酬を，ポジションを解消した時に一括で決定し，付与するよう変更を加えた. 以上のことを踏まえた, NT 倍率取引における投資戦略を，深層強化学習によって獲得する数理モデルを構築し，その有用性を確認した.

2. 提案手法

本研究はコンピュータシミュレーションによって行う．コンピュータ上につくられた仮想的な投資家が，1 日 1 回市場の状態を観測し，その状態におけるそれぞれの投資行動の価値（Q 値）を推測する．その価値が高い行動を選択，実行し，結果が良ければその行動に報酬を与えて，同じ状態においてその

行動をとりやすくする．この Q 値の推測はニューラル・ネットワークを用いて行い，報酬に応じてその重みを変えることを繰り返して学習を進めていく．

2.1 既存手法からの変更点

本研究では，松井らの手法[３]をベースに総資産の最大化を目的として，以下の点を変更した．

(1) 取引手法

松井らの手法では，日本国債の週次取引に対する行動規則を学習した．しかし，国債には多くの価格変動要因が存在し，適切な行動選択を困難にしている．これらの変動要因をすべて取り入れて行動を選択することは不可能である上，多くの場合取り入れていない要因からも大きな影響を受けるため，安定した学習ができなくなってしまう．そこで，まず「考慮しなければならない価格変動要因を減らし，状況を簡略化すること」を考えた．具体的には，相関性が強く，価格差が拡大しても元に戻りやすいような 2 つの金融商品に対して，「買い」と「売り」をそれぞれ同時に行う取引を考える．これにより価格変動要因の大部分を相殺可能である．このような相関が強い金融商品として，日経 225 先物と TOPIX 先物がある．この 2 銘柄に対して「買い」と「売り」をそれぞれ同時に行う取引を NT 倍率取引という．日経 225 先物と TOPIX 先物の価格の推移を図 1 に示す．

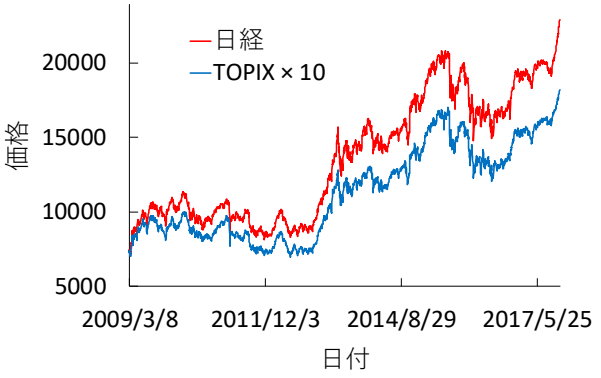

図 1：日経 225 先物と TOPIX 先物の価格推移．

図 1 の横軸は期間，縦軸は価格である．赤い折れ線が日経 225 先物であり，青い折れ線が TOPIX 先物である．日経平均株価と TOPIX には約 10 倍の違いがあるため，この図では TOPIX に 10 をかけたものをプロットしている．これを見ると，変動の仕方がかなり似通っていることが分かる．これは，日経平均株価と TOPIX がどちらも東証一部上場企業の株価や時価総額から計算される指標だからであり，変動の仕方がわずかに異なるのは計算に用いられている企業や，株価か時価総額かの違いによるものである．このように，定量化が困難な各国のニュースなどの影響の大部分はどちらも等しく受けており，2 銘柄の価格の違いに着目した投資判断を行うことによって，価格変動要因の大部分が相殺された状態での取引が可能になる．そこで本研究では，NT 倍率取引を取引手法として選択した．

(2) 学習方法

松井らの手法では，取引量を調節しながら利益率の複利効果を最大化するため，投資比率と複利リターン[２]を考慮した学習を行っている．しかし，本研究ではモデルを単純化するため，取引を 1 単位ずつの売買もしくはポジションの解消に制限した．

(3) 行動

本研究では行動として「1 単位 NT 買い（日経 225 先物買い，TOPIX 先物売り）」，「1 単位 NT 売り（TOPIX 先物買い，日経 225 先物売り）」，「NT 買いポジション解消」，「NT 売りポジション解消」，「何もしない」の 5 つとする．ここで，日経 225 先物の最低取引単位（1 単位）は日経平均株価の 1,000 倍，TOPIX 先物の最低取引単位（1 単位）は TOPIX の 10,000 倍である．NT 買い（売り）ポジションとは，日経 225 先物を 1 単位以上保持（空売り），TOPIX 先物を 1 単位以上空売り（保持）している状態を指し，それを解消することはすべての金融商品を現金化することを指す．

(4) 状態

松井らの手法では，状態変数として終値とその標準偏差を相対化した値を用いている．これは金融商品の価格などは大きく変動するため，そのまま状態として用いると，体験したことのない未知の状態に陥りやすくなってしまうからである． 時刻 t の状態変数 v_t を相対化した値 O_t は以下のように求める．

$$O_t = \frac{v_t - \mu_{t,k}}{4\sigma_{t,k}} \tag{1}$$

ここで，$\mu_{t,k}$ は時刻 t から過去 k 期間のデータから求めた移動平均，$\sigma_{t,k}$ は同様に求めた移動標準偏差を表す．これにより，$\left[\mu_{t,k} - 4\sigma_{t,k},\ \mu_{t,k} + 4\sigma_{t,k}\right]$ の範囲を $[-1, 1]$ の範囲に正規化できる．松井らは終値とその移動標準偏差をそれぞれ相対化した 2 つの状態変数を用いている．

本研究では，深層強化学習の多数の状態変数を扱えるという利点を活かし，より状況を適切に捉えるため，状態変数の数を 10 に増やす．まず，TOPIX 先物の終値に対する日経 225 先物の終値の

割合である NT 倍率と，その移動標準偏差を相対化した値を状態変数とした．この時，移動平均を求める期間 k は短期，中期，長期の 3 パターン設定し，それぞれに対して相対化を行う．NT 倍率は，松井らの終値と同様に現在の市場の動向を表す指標として採用している．次に利益確定を学習するために「含み損益」を加えた．t 日目の含み損益 $prof_t$ は以下のように定義する．

$$prof_t = \frac{(P_t^N - P_{t-e}^N)S_t^N + (P_t^T - P_{t-e}^T)S_t^T}{A_0} \quad (2)$$

ここで，P_t^N は t 日目の N（日経 225 先物）の価格，P_t^T は t 日目の T（TOPIX 先物）の価格，e はポジションをとってからの日数である．よって，P_{t-e}^N はポジションをとった時の価格になる．S_t^N は t 日目の N（日経 225 先物）のストック数であり，保有している分を正の値，空売りしている分を負の値で表す．A_0 は初期資産である．これを状態変数として取り入れることで，今ポジションを解消したらどのくらい利益が得られるかを把握することができる．次に「“NT 買いポジションをとってからの最大 NT 倍率”と“現在の NT 倍率”の差」と「“現在の NT 倍率”と“NT 売りポジションをとってからの最低 NT 倍率”の差」をそれぞれ「機会損失幅（NT 買いポジション）」と「機会損失幅（NT 売りポジション）」として定義し，状態変数として導入する．これらは，最大利益を獲得できる時点から NT 倍率がどのくらい変わってしまったかを把握するための状態変数である．そして，現在のポジションを把握するための「現在のポジション」を加えた 10 個の状態変数を用いて学習を行う．

(5) 報酬

松井らの手法では，複利リターンを最大化するため，利益率 R，投資比率 f の時のグロス利益率（利益率に 1 を加えたもの，つまりは資産の変化前に対する変化後の割合である）の対数 $log(1 + Rf)$ を報酬としている．しかし，本研究では複利リターンを考慮しない．

また，松井らの手法ではとった行動に対してすぐに報酬を決めて与えているが，金融取引において行動の良し悪しをすぐに決めるのは大変困難である．そこで本研究では，ポジションを取得してから解消するまでの全ての行動に対する報酬を，ポジションを解消した後に一括で決定し，付与する．このとき付与量はポジションの状態によって異なるように設定した．買い（売り）ポジションの取得時と保持時には，「“最大（最低）NT 倍率”と“現在の NT 倍率”の差の絶対値」を報酬とする．このとき，最大（最低）NT 倍率の時点より前の行動に関してはそのままで正の報酬，後の行動に関しては -1 をかけて負の報酬とする．これにより，前者は「現在の NT 倍率からこのポジション中に NT 倍率がどれだけ上がる（下がる）か」，後者は「最大で稼げる NT 倍率からどのくらい下がって（上がって）しまったか」を考慮した報酬を表す．買い（売り）ポジションの解消時には「“現在の NT 倍率”と“ポジション取得時の NT 倍率”の差（“ポジション取得時の NT 倍率”と“現在の NT 倍率”の差）」を報酬とする．これは「ポジションを取得した時の NT 倍率からどれだけ上がった（下がった）か」，つまり利益をどれだけ出せたかに応じた報酬であることを表す．

さらに，持っていないポジションを解消しようとした際に定数 $penalty$ $(penalty < 0)$ の報酬を与える．例えば，NT 買いポジションをとっている時に NT 売りポジションを解消しようとした時などである．このような行動をとらないように負の報酬を設定した．

2.2 提案手法の流れ

提案手法での学習の流れを以下で述べる．

① 初期化

行動価値関数を表すニューラル・ネットワークを初期化する．

② 取引とデータ収集

行動価値関数から得られる行動規則に従って取引を行い，データ（状態変数ベクトル X，行動 a，報酬 r，次の状態を表す状態変数ベクトル X'）を収集する．収集したデータは $Replay\ Buffer$ に保存するが，この時，ポジションの状態に応じて異なる処理を行う．ポジションを保持していないときに「何もしない」を選択した場合は報酬 $r = 0$ とし，得られたデータを即座に $Replay\ Buffer$ に加える．ポジションを取得した時から解消する時までのデータは即座には報酬を決めずに，一旦 $Temp\ List$ に保存する．これらのデータは，ポジションを解消した時に報酬をまとめて決定し，$Replay\ Buffer$ に保存する．その後，$Temp\ List$ 内のデータをすべて削除する．

この際の行動選択には，ε の確率でランダムに行動し，それ以外は Q 値の一番高い行動を選択する $\varepsilon-$greedy 法を用いる．これを M 回繰り返す．

③ ニューラル・ネットワークの更新

$Replay\ Buffer$ 内のデータからランダムサンプリングにより，m 個取り出してそれぞれ Q 値を計算し，それらを教師データとして行動価値関数を表すニューラル・ネットワークを更新する．ここで，

t 日目の状態 X での行動 a に対する Q 値, つまり, X を入力した時の望ましい出力 q_t は以下のように求める.

$$q_t \leftarrow r + \gamma \max_{a'} Q(\mathrm{X}', a') \quad (3)$$

ここで, r は 2.1 で決めた報酬, γ は将来の報酬に対する割引率である. これにより, 今回の行動で得られた報酬と, 次の状態での最大価値を持つ行動の Q 値を割り引いたものの和を望ましい出力とする.

④ 終了判定

②～③を任意の回数繰り返す.

テスト時には, 行動価値関数から得られる行動規則に従い, テスト期間の取引を行う. この際, 行動選択には, 常に Q 値の一番高い行動を選択する greedy 法を用いる.

3. 実験

実験は日経 225 先物と TOPIX 先物の日次取引を対象として行う. 学習期間は 2009/3/4～2015/12/31 で 1682 日分, テスト期間は 2016/1/4～2017/12/29 で 506 日分のデータを用いた. また, 過去の実験より, NT 倍率が長期的に上昇トレンドであることから, 単に「1 単位 NT 買い（日経 225 先物買い, TOPIX 先物売り）をし続けてしまう」という局所解に陥ってしまったという問題があった. そこで本実験では, 学習期間の NT 倍率の時系列データからトレンドを除去する. 原系列とトレンド除去後の NT 倍率の推移を図 2 に示す.

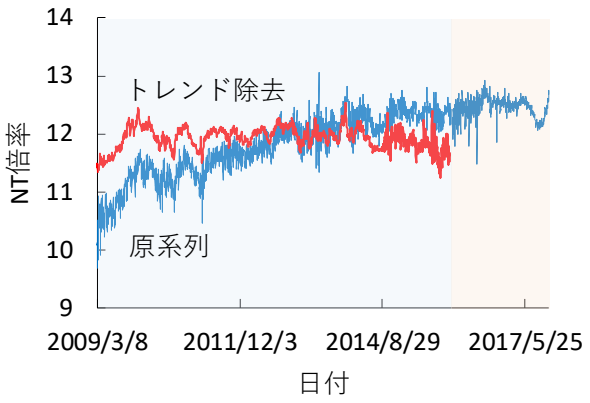

図 2：NT 倍率の推移.

図 2 の横軸は日付, 縦軸は NT 倍率である. 青い折れ線は NT 倍率の原系列, 赤い折れ線はトレンド除去後の NT 倍率を表している. 青背景の部分が学習期間, 橙背景の部分がテスト期間である. 取引は 1 日 1 回, 前日の終値を観測し, 当日の始値で行う. 学習期間での取引をすべて終えるまでを 1 エピソードと定義し, 1000 エピソードを終える度にテスト期間の取引を行い, それを終えたらまた学習期間の取引を行う.

本研究で用いる深層強化学習のモデルは Deep Q-Network である. ここで用いられるニューラル・ネットワークの中間層は 2 つで, そのユニット数は入力側から 100, 25 である. 重みは Xavier の初期値を用い, 活性化関数は, 中間層から出力層の間が線形結合, それ以外はランプ関数（ReLU）とした. 最適化手法は Adam, 学習時のニューラル・ネットワークの更新間隔は $M = 100$, ランダムサンプリング数は $m = 20$ である.

学習期間の行動選択方法は ε－greedy 選択, テスト期間は greedy 選択とした. ランダムな行動を選ぶ確率 ε は 0 エピソード時には 1.00 とし, 50,000 エピソードかけて 0.10 まで線形に低下していくように設定した. Q 値更新時の将来報酬の割引率は $\gamma = 0.95$ とした.

状態変数の相対化に用いる期間は $k = 5, 25, 75$, 違反行動をとった時の報酬は $penalty = -0.1$ とし, 初期資産は $A_0 = 10{,}000{,}000$ で実験を行った.

4. 結果と考察

まず, 学習期間の最終総資産の推移を図 3 に示す.

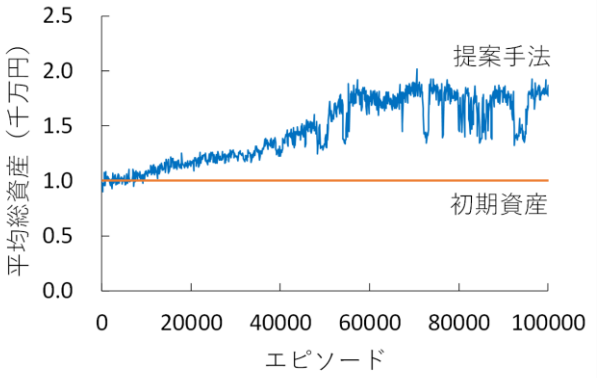

図 3：学習期間の平均最終総資産の推移.

図 3 は横軸がエピソード, 縦軸が総資産である. 青い折れ線は, 1 エピソードの終わり時点での総資産を 100 エピソード毎に平均し, プロットしたものである. また, 橙色の直線は初期資産である. これを見ると提案手法は, ゆるやかに総資産を伸ばし, 最終的にはかなり高い値で収束していることが分かる.

次に，テスト期間の最終総資産の推移を図 4 に示す．

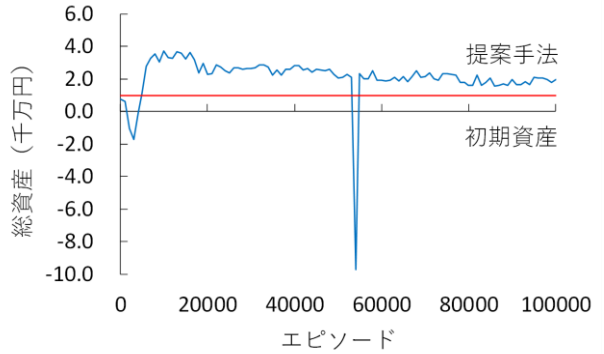

図 4：テスト期間の最終総資産の推移．

図 4 は横軸がエピソード，縦軸が総資産である．青い折れ線は，テスト期間の取引結果の最終総資産をプロットしたものである．また，赤色の直線は初期資産である．提案手法による最終総資産は，序盤落ち込んだ後，急速に上昇し，その後は徐々に落ちていく様子が観測された．また，中盤には急激に最終総資産が落ち込んでいるが，この原因はまだ分かっていない．

次に，本実験における最終エピソードである 10 万エピソード時にどんな戦略をとっているのかについて分析する．まずはエピソード中の総資産の推移を確認する．10 万エピソード時のテスト期間 1 試行中の総資産の推移を図 5 に示す．

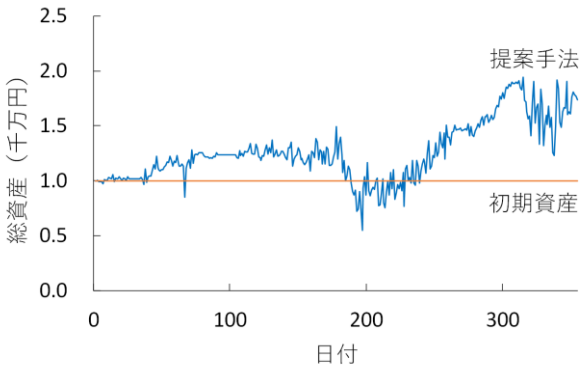

図 5：テスト期間 1 試行中の総資産の推移．

図 5 は横軸が日付，縦軸が総資産である．青い折れ線は，10 万エピソード時のテスト期間 1 試行中の総資産を 1 日毎にプロットしたものである．また，橙色の直線は初期資産である．これを見ると，最終的には初期資産よりも高い値で取引を終えているが，中盤に初期資産を下回る部分がいくつか存在しており，安定して稼げているとは言い難い．

次に，どのような取引を行っているかを先程同様 10 万エピソード時のテスト期間 1 試行の結果を用いて分析する．具体的には，NT 買いストック数（日経ストック数）と NT 倍率の推移を見て，NT 倍率がどうなっている時にどのようなポジションをとっているのかを確認する．図 6 は NT 買いストック数と NT 倍率の推移である．

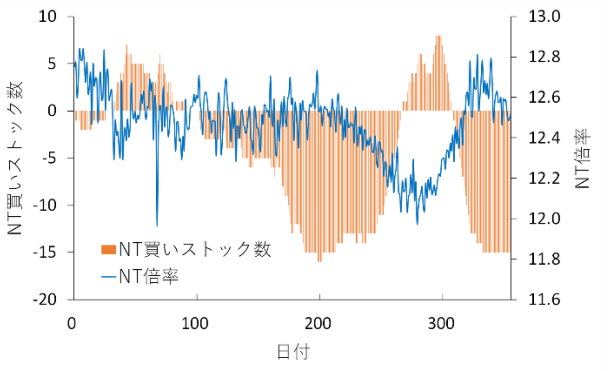

図 6：NT 買いストック数と NT 倍率の関係．

図 6 は横軸が日付，左軸が NT 買いストック数，右軸が NT 倍率である．橙色の棒グラフが NT 買いストック数，青色の折れ線が NT 倍率である．NT 買いストック数が正のときは，NT 買いポジションをとっていることを表し，NT 買いストック数が負のときは，NT 売りポジションをとっていることを表す．振る舞いを見ると，NT 倍率が下がった時には NT 買いストック数は負の値になっており，逆に NT 倍率が上がった時には NT 買いストック数は正の値になった．このような「価格が上がると思われるときに買い，上がりきったところで売る（価格が下がると思われるときに売り，下がりきったところで買う）」戦略は「順張り」といい，金融取引で一般的に使われる戦略である．これは，報酬により，ポジション取得時点から NT 倍率がどのくらい変化するか（どのくらい利益を出せるか）を考慮していることが効いていると考えられる．これによって「『上がると思われるときに買いポジションをとる』または，『下がると思われるときに売りポジションをとる』」という行動規則が創出されている．また，状態変数の「機会損失幅」があることにより，「一番含み益を出している時点から NT 倍率がどれくらい変化しているか」の報酬を，状態をしっかり把握させた状態で与えることができている．これにより，「『買いポジションをとっているときに NT 倍率が下がりそうになったら買いポジションを解消する』または，『売りポジションをとっているときに NT 倍率が上がりそうになったら売りポジションを解消する』」という利益確定

の動きが創出できている．

しかし，行動の中に「NT 買いポジション解消」，「NT 売りポジション解消」があるにも関わらず, NT 買いストック数は瞬時に 0 になることはなく，1 単位ずつ変化している．これはポジションを解消すべきタイミングで「NT 買い」または「NT 売り」の行動をとって1単位ずつポジションを崩しており,「NT 買いポジション解消」,「NT 売りポジション解消」の行動を上手く使えていないことを表している．原因としては，1 つずつポジションを崩してポジションを解消した場合と，一気に解消した場合で報酬量が変わらないことが考えられる．

5. 今後の課題

まず,「買いポジション解消」と「売りポジションの解消」の 2 つの行動を有効に活用できるようにすることが挙げられる．ポジションを 1 単位ずつ崩していくのではなく一気に解消することで，素早い取引が可能になると考えられる．また，テスト期間の結果 (図 4) の総資産が急落している問題についての分析を行う必要がある．

そして最後に，最新手法の利用を検討している．現在は Deep Q-Network という学習方法を用いているが，A3C (Asynchronous Advantage Actor-Critic) [４] という Deep Q-Network を発展させたモデルが開発されている．このモデルには，Asynchronous（複数のエージェントを同時に動かし，個々の経験を集めて学習），Advantage（1 ステップ先ではなく，数ステップ先の報酬を考慮）などの特徴がある．これを用いることで学習時に 1 つ先の報酬だけでなく，もう少し先の報酬も考慮できるようになる他，LSTM[５]などの時系列データの扱いに長けたニューラル・ネットワークの使用が可能になる．このような理由から，A3C の導入を検討している．

参考文献

[１] Richard S. Sutton and Andrew G. Barto：強化学習（三上貞芳・皆川雅章 訳）．森北出版(2000)

[２] 松井藤五郎, 後藤卓, 和泉潔, 陳ユ：複利型強化学習における投資比率の最適化，人工知能学会論文誌, Vol.28, No.3, pp. 267-272 (2013)

[３] 松井藤五郎, 片桐雅浩:金融取引戦略獲得のための複利型深層強化学習，第 16 回人工知能学会金融情報学研究会(SIG-FIN), SIG-FIN-016-01 (2016)

[４] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, Koray Kavukcuoglu : Asynchronous Methods for Deep Reinforcement Learning, In Proceedings of the 33rd International Conference on Machine Learning (ICML), pp. 1928–1937 (2016)

[５] Sepp Hochreiter, Jürgen Schmidhuber : Long Short-Term Memory, Neural computation, 9(8), pp. 1735–1780 (1997)