論　文

多期間消費投資モデルにおける強化学習を用いたポートフォリオ戦略

奥原　浩之[†]　　柴田　淳子[††]　　田中稔次朗[†]　　坂和　正敏[††]

Portfolio Selection by Reinforcement Learning in Multiple Period Consumption and Investment Model

Koji OKUHARA[†], Junko SHIBATA[††], Toshijiro TANAKA[†], and Masatoshi SAKAWA[††]

あらまし　本研究では，人工市場において構築された多期間消費投資モデルにおけるポートフォリオ戦略がどのような特性をもつかを分析する．市場には無危険資産と危険資産が存在する．投資家を模倣するエージェントは各自の判断に基づいて，消費に関する期待効用を最大化するべく自己の資産の再配分を決定する．この投資家の合理的な動作を強化学習をエージェントに適用することで仮想的にコンピュータ上で実現する．その上で，エージェントのもつ将来の報酬に対する割引や効用関数に対する選好の程度が形成される金融価格や消費に関する期待効用へ及ぼす影響について分析する．

キーワード　人工市場，多期間消費投資モデル，エージェント，強化学習，ポートフォリオ戦略

1. まえがき

ファイナンス工学は過去 30 年間に急速に発展した分野であり，近年の金融市場における規制緩和や国際化を理論的な観点から支援している．更に情報技術の進展に伴い，多くの金融商品を組み込んだ投資戦略や資産運用のためにファイナンス理論の適用が可能となり，金融市場を取り巻く状況も大きく変化してきている [1]．ファイナンス工学の主な課題の一つに資産選択問題がある．資産選択問題とは危険資産と無危険資産からなる市場において，投資家が消費に関する期待効用を最大化するべく自己の資産の再配分を決定するものである．

従来のファイナンス理論における資産選択問題に関する多くの数理モデルは 1 期間を対象としている．しかしながら，実際の運用は資金の流入，流出，リバランスが繰り返されていることから，運用期間に長短がある多期間であると考えられる．多期間資産選択問題の厳密な最適ポートフォリオ戦略を求めることは一般に困難であるため，完全で効率的な市場における無裁定条件を仮定した近似モデルに対してポートフォリオ戦略が求められている [2]．ところが現実の市場において，これらの仮定が常に成り立っていることを確認することも困難である．

効率的市場仮説に基づいたファイナンス理論では，金融価格の振舞いは確率過程で表現される．しかし現実の市場では，投資家は利用可能な情報を迅速かつすべて適切に反映して行動しているとは限らない．そこで，金融価格が投資家の行動の結果として形成されるマルチエージェントによる人工市場の研究が盛んに行われる傾向にある [3]．

このように金融市場の環境と投資家の行動による複雑な相互作用が存在する状況において，投資家がその環境をあらかじめ詳細に把握し，それに基づいて行動を決定することができない場合には，環境との相互作用に基づいて試行錯誤的に行動規範を獲得する手法が有効となることが考えられる．そのような手法の一つに強化学習がある．そこで本研究ではマルチエージェントによる多期間消費投資モデルを考え，各エージェントが強化学習 [4] を利用してポートフォリオ戦略を導出する人工市場を構築し，エージェントのもつ将来の報酬に対する割引や効用関数に対する選好の程度が形成される金融価格や消費に関する期待効用へ及ぼす

† 広島県立大学経営学部経営情報学科，庄原市
Department of Management and Information Sciences, Hiroshima Prefectural University, Shobara-shi, 727–0023 Japan
†† 広島大学大学院工学研究科複雑システム工学専攻，東広島市
Department of Artificial Complex Systems Engineering, Graduate School of Engineering Hiroshima University, Higashi-hiroshima-shi, 739–8527 Japan

影響について分析する．

2. 多期間消費投資モデルの概要

ここで，多期間消費投資モデルの概要について述べる [5]．市場には無危険資産 ($i = 1$) と危険資産 ($2 \le i \le M$) が存在する．資本市場は一般に完全であると仮定し，

仮定 1) 取引手数料，配当及び税金がない．
仮定 2) 投資家の投資量は任意の実数である．
仮定 3) 投資家は価格及び収益に影響を与えない．
仮定 4) 投資家は空売りにより収益を達成できる．

とする．

金融市場は投資により収益が上げられることを保障するために，投資収益について以下の基本的な性質を仮定する．

$$\begin{aligned} &r_{1t} \ge 0 && (t = 1, 2, \cdots) \\ &\mathrm{E}[r_{it}] \ge \delta + r_{1t} && (\delta \ge 0,\ \exists i, t) \\ &\mathrm{E}[r_{it}] \le K && (\forall i, t) \end{aligned}$$

ここで，r_{1t} は t 期の利子率，r_{it} は t 期に投資機会 i ($i = 2, \cdots, M_t$) について資本 1 単位から得られる収益（確率変数）を表す．もし期初に第 i 機会に θ を投資した場合，期末には $(1 + \gamma_{it})\theta$ を得る．M_t は t 期に利用可能な投資機会数 ($M_t \le M$) である．

また，（非定常）収益率分布 F_t

$$\begin{aligned} &F_t(z_2, z_3, \cdots, z_{M_t}) \\ &\quad \equiv \Pr\{r_{2t} \le z_2, r_{3t} \le z_3, \cdots, r_{M_t t} \le z_{M_t}\} \end{aligned}$$

は各期において独立，あるいはマルコフ連鎖に従うと仮定する．任意の t，及び空売りが不可能な投資機会 i について $\theta_i \ge 0$ かつ $\sum_{i=2}^{M_t} \mid \theta_i \mid = 1$ である任意の θ について収益が負となる可能性の存在条件

$$\Pr\left\{\sum_{i=2}^{M_t}(\gamma_{it} - \gamma_{1t})\theta_i < \delta_1\right\} > \delta_2$$

が満たされるとする．ここで，$\delta_1 < 0$，$\delta_2 > 0$ である．上式は無裁定条件と等価であり，ポートフォリオ問題が解をもつための必要十分条件である．

次に投資家は各期に支払能力を維持していなくてはならないため，意思決定時点 t（t 期末）における資本の投入量 w_t において支払可能制約が満たされるものとする

$$\Pr\{w_t \ge 0\} = 1 \qquad (t = 1, 2, 3, \cdots, T-1)$$

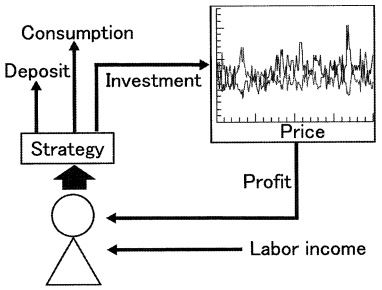

図 1 動的消費投資モデル
Fig. 1 Dynamical consume and investment model.

$t-1$ 期に投資された資本は

$$\sum_{i=1}^{M_t} z_{it} = w_{t-1} - c_t \tag{1}$$

である．ここで，z_{1t} は t 期の貸出金量（$z_{1t} < 0$ は借入れを示す），z_{it} は t 期初での投資機会 i ($i = 2, \cdots, M_t$) への投資量，c_t は t 期初に消費に振り向けられる量である．

消費と労働収入の問題を含む基本的な動的消費投資モデルにおいて（図 1 参照），時刻 t の投資価値を無危険資産，危険資産に分解して表すならば，

$$\begin{aligned} w_t = &\sum_{i=2}^{M_t}(1 + r_{it})z_{it} + (1 + r_{1t}) \\ &\times \left(w_{t-1} - \sum_{i=2}^{M_t} z_{it} - c_t\right) + y_t \end{aligned} \tag{2}$$

である．ここで，y_t は t 期末に受け取る労働収入である．これより以下の基本差分方程式を得る．

$$\begin{aligned} w_t = &\sum_{i=2}^{M_t}(r_{it} - r_{1t})z_{it} + (1 + r_{1t})(w_{t-1} - c_t) + y_t \\ &(t = 1, 2, 3, \cdots, T) \end{aligned}$$

投資家の目的は，$c_t \ge 0$ のもとで消費系列からの期待効用を最大化すること

$$\mathrm{Max}\ \mathrm{E}[U(c_1, \cdots, c_T)]$$

である．ここで，効用関数 U は単調増加，狭義凹，及び投資家の選好

$$(a, b, c_3, \cdots, c_T) \succeq (b, a, c_3, \cdots, c_T), \qquad (a > b)$$

を反映するものとする．更に，以下のような強い仮定を考慮する．

仮定 5)　個人の寿命（計画期間）は認知である．
仮定 6)　利子率が決定論的である．
仮定 7)　労働収入は決定論的である．
仮定 8)　効用関数が時間加法的である．
したがって，仮定 7) より

$$Y_{t-1} \equiv \frac{y_t}{r_{1t}} + \cdots + \frac{y_T}{(1+r_{1t})\cdots(1+r_{1T})}$$

となり，仮定8) よりすべての t について $U_t^{'} > 0$, $U_t^{"} < 0$，そして通常 $\alpha < 1$ として，

$$U(c_1, \cdots, c_T) = u_1(c_1) + \alpha u_2(c_2) + \cdots + \alpha^{T-1} u_T(c_T) \qquad (3)$$

となる．ここで，期間 t における効用関数は

$$u_t(c_t) = \frac{1}{\gamma} c_t^{\gamma} \qquad (0 < \gamma < 1)$$

であり，γ は効用関数の選好の程度を表しており，あるものを別のものよりも好んで選ぶ度合を反映している．γ が大きくなると消費が大きくても効用が飽和することなくほぼ線形に増加し，小さくなるとある程度までの消費については効用が大きく変化するが，それを超えると効用が飽和していく．

3. 人工市場の概要と強化学習によるポートフォリオ戦略

3.1 人工市場の概要

ここでは，先のモデルのいくつかの条件を緩和することにより得られる，より現実に近い人工市場の枠組みについて述べる．状態は離散時間で推移するとし，$t-1$ から t の間を t 期とする．また，市場には無危険資産 $(i=1)$ と危険資産 $(2 \le i \le M)$ が存在するものとする．更に，投資者を表すエージェントが N 人存在するものとする．エージェントは均一な振舞いをする必要はない．

t 期におけるエージェント n の資産 i への投資量を z_{it}^n とし，消費に振り向けられる量を c_t^n とする．人工市場においては，企業は自身の銘柄を保持しているエージェントに配当金 d_{it} を配当する．そのため，時刻 t における投資価値である式 (2) は

$$w_t^n = \sum_{i=2}^{M_t^n} (1+r_{it}) z_{it}^n + (1+r_{1t}) z_{1t}^n + y_t^n + \sum_{i=2}^{M_t^n} \left(r_d d_{it} \frac{z_{it}^n}{p_{i(t)}} + z_{it}^{'n} \right)$$

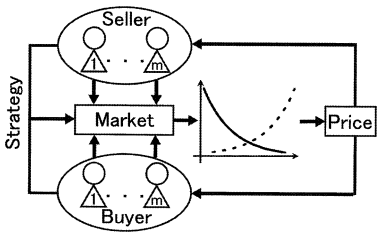

図 2　人工市場モデル
Fig. 2　Artificial market model.

となる．ここで，r_d は正の定数，$z_{it}^{'n}$ は取引不成立の場合の返金である．収益 r_{it} は

$$r_{it} = \frac{p_{i(t)}}{p_{i(t-1)}} - 1$$

で与えられる．

配当は離散の有色ノイズであるとして，

$$\log \frac{d_{it}}{\bar{d}_{it}} = \epsilon_i^a \log \frac{d_{i(t-1)}}{\bar{d}_{it}} + \epsilon_i^b \xi_{it}$$

で与えることとする [3]．ここで，$\xi_{it}(t)$ は平均 0，分散 σ_i^2 のガウスノイズであり，ϵ_i^a, ϵ_i^b は $(\epsilon_i^a)^2 + (\epsilon_i^b)^2 = 1$ を満たす正のパラメータである．$\log d_{it}/\bar{d}_{it}$ は平均 0，分散 σ_i^2 で，相関時間 $\tau_s = 1/\log(\epsilon_i^a)$ で自己相関関数が減衰する．

各エージェントは次の期間が始まると同時に保持している資産の量を変化させたり，消費を行うことができる（図 2 参照）．ただし，ここでは式 (1) と同様に全資産が一定のもとで組替えを行い

$$\sum_{i=1}^{M_t^n} z_{it}^n = w_{t-1}^n - c_t^n$$

市場における危険資産の量はそれぞれ一定

$$\sum_{n=1}^{N} z_{it}^n = Z_i \qquad (2 \le i \le M)$$

であるとする．

各エージェントは有限の資産のもとでリスクを考慮した各自の判断に従い式 (3) の期間 T を十分大きく考えた効用を最大にするべく行動するものとする．ところで，各エージェントは常に希望どおりの売買ができるとは限らない．今，エージェント n が希望する危険資産 i の売買量を b_{it}^n, o_{it}^n とすると，危険資産 i に関する全エージェントが希望する売買量は

$$B_{it} = \sum_{n=1}^{N} b_{it}^{n}$$

$$O_{it} = \sum_{n=1}^{N} o_{it}^{n}$$

となる．そのため，$B_{it} = O_{it}$ の場合のみ希望どおりの売買が可能であり，$B_{it} \neq O_{it}$ の場合における各エージェントの危険資産の保有量は

$$z_{it}^{n} = z_{i(t-1)}^{n} + \frac{V_{it}}{B_{it}} b_{it}^{n} - \frac{V_{it}}{O_{it}} o_{it}^{n}$$

となるものとする．ここで，$V_{it} \equiv \min(B_{it}, O_{it})$ である．そこで，取引不成立の場合の返金 $z_{it}^{'n}$ は

$$z_{it}^{'n} = \left(1 - \frac{V_{it}}{B_{it}}\right) b_{it}^{n} - \left(1 - \frac{V_{it}}{O_{it}}\right) o_{it}^{n}$$

となる．

危険資産 i の価格 p_{it} は主にすべてのエージェントの売買に基づいて決定される．そのような状況下で，価格 p_{it} を次のように決定する．

$$p_{i(t+1)} = \frac{2p_{it}}{1 + \exp\{-U_{it}/T_i^P\}}$$

$$\text{ただし } U_{it} = \log \frac{B_{it}}{O_{it}}$$

ここで，T_i^P は危険資産 i の感度を表す正の定数であり，小さいと需要と供給の差に敏感であり，大きいと鈍感であることを表す．この更新式は文献 [3] の価格の更新式に基づいている．

以上のように，本研究で対象とする人工市場は無限期間に及ぶ消費系列からの期待効用の最大化を目的とし，従来の仮定 1)，3)，5) とは異なり，配当が存在し，エージェントの行動が価格及び収益に影響を与え，エージェントの計画期間は必ずしも認知される必要がないと考えることが可能であるなど多期間消費投資モデルとは大きく異なるといえる．

3.2 強化学習によるポートフォリオ戦略

人工市場において，時刻 t にエージェント n が有限の資産のもとでリスクを考慮して各自の判断に従い行動するときの利得を

$$V_t^n = \sum_{k=0}^{\infty} (\alpha_n)^k u_{t+k}(c_{t+k}^n)$$

とする．ここで，α_n はエージェント n のもつ割引率である．マルコフ決定過程においてエージェントが定常政策 π をとる場合，利得の期待値は状態価値関数といい，以下の Bellman 方程式を満たす [7]．

$$V_\pi^n(s) = \mathrm{E}_\pi \left[\sum_{k=0}^{\infty} (\alpha_n)^k u_{t+k}(c_{t+k}^n) | s_t^n = s \right]$$

$$= \sum_a \pi(s,a) \sum_{s'} P_{ss'}^a \{ R_{ss'}^a + \alpha_n V_\pi^n(s') \}$$

ここで，政策 $\pi(s,a)$ は状態 $s \in S$ において行動 $a \in A(s)$ をとる確率，$P_{ss'}^a$ は政策 $\pi(s,a)$ で状態が s から s' へ遷移する確率，$R_{ss'}^a$ は政策 $\pi(s,a)$ で状態が s から s' へ遷移したときの期待利得である．また，行動価値関数として

$$Q_\pi^n(s,a) = \mathrm{E}_\pi \left[\sum_{k=0}^{\infty} (\alpha_n)^k u_t(c_t^n) | s_t^n = s, a_t^n = a \right]$$

$$= \pi(s,a) \sum_{s'} P_{ss'}^a \{ R_{ss'}^a + \alpha_n V_\pi^n(s') \}$$

を定義する．

マルコフ決定過程においては，ほかの政策より同等か優れた政策が少なくとも一つは存在し，これを最適政策 π^* という．最適政策 π^* に対する状態価値関数は

$$V_*^n(s) = \operatorname*{Max}_a V_\pi^n(s) \qquad (\forall s \in S)$$

行動価値関数は

$$Q_*^n(s,a) = \operatorname*{Max}_\pi Q_\pi^n(s,a) \qquad (\forall s \in S,\ \forall a \in A)$$

と定義される．

最適な価値関数に対する Bellman 方程式である Bellman 最適方程式は

$$V_*^n(s) = \operatorname*{Max}_a Q_*^n(s,a)$$

$$= \operatorname*{Max}_a \sum_{s'} P_{ss'}^a \{ R_{ss'}^a + \alpha_n V_\pi^n(s') \}$$

となる．この Bellman 最適方程式を近似的に解くための手法の一つに強化学習がある．

そこで，本研究では，人工市場におけるポートフォリオ戦略をニューラルネットを用いた強化学習により求めることを提案する．強化学習には Actor-Critic モデル [8] を実現するニューラルネットワーク [9] を適用する．

まず，エージェント n は環境において状態 $\mathbf{x}_t^n$ を観測し，Actor は制御出力を

$$q_{it}^n = f\left(\sum_{j=1}^{N_A} W_{ijt}^{A_n} g_j^{A_n}(\mathbf{x}_t^n) + rnd_{it}\right)$$

$$g_j^{A_n}(\mathbf{x}_t^n) = \exp\left\{-\frac{1}{2}(\mathbf{x}_t^n - \mathbf{m}_j^{A_n})^{\mathrm{T}} \mathbf{C}_j^{A_n{}^{-1}} \times(\mathbf{x}_t^n - \mathbf{m}_j^{A_n})\right\}$$

で生成する．ここで，g_j^A は j 番目の動径基底関数，N_A は Actor の動径基底関数の数，$W_{ijt}^{A_n}$ は結合荷重，rnd_{jt} は正規乱数であり，f はシグモイド関数

$$f(x) = \frac{1}{1+\exp\{-x/T_n^A\}}$$

である．ここで，T_n^A はエージェント n の感度である．

Actor の制御出力からエージェント n は

$$b_{it}^n = q'^n_{it} w_t^n - z_{i(t-1)}^n \qquad (q'^n_{it} w_t^n > z_{i(t-1)}^n)$$
$$o_{it}^n = z_{i(t-1)}^n - q'^n_{it} w_t^n \qquad (q'^n_{it} w_t^n < z_{i(t-1)}^n)$$

で取引を行おうとする．ここで，q'^n_{it} は以下で定義される．

$$q'^n_{it} = \frac{q_{it}^n}{\sum_{i=1}^N q_{it}^n}$$

Critic は評価値

$$V_\pi^n(\mathbf{x}_t^n) = \sum_{j=1}^{N_C} W_{jt}^{C_n} g_j^{C_n}(\mathbf{x}_t^n)$$

を生成する．ここで，N_C は Critic の動径基底関数の数であり，動径基底関数は Actor で用いたものと同様に

$$g_j^{C_n}(\mathbf{x}_t^n) = \exp\left\{-\frac{1}{2}(\mathbf{x}_t^n - \mathbf{m}_j^{C_n})^{\mathrm{T}} \mathbf{C}_j^{C_n{}^{-1}} \times(\mathbf{x}_t^n - \mathbf{m}_j^{C_n})\right\}$$

で与えた．

行動の結果，Critic は環境から報酬

$$reward_t^n = u_t(c_t^n) = \frac{1}{\gamma_n}\left(w_{t-1}^n - \sum_{i=1}^{M_t^n} z_{it}^n\right)^{\gamma_n}$$

を受け取り，遷移後の状態 $\mathbf{x}_{t+1}^n$ を観測する．更に，強化信号として，時刻 t における期待効用

$$\mathrm{E}[u_t(c_t^n)] = V_\pi^n(\mathbf{x}_t^n) - \alpha_n V_\pi^n(\mathbf{x}_{t+1}^n)$$

と実際の効用との差

$$\begin{aligned}\delta_t &\equiv u_t(c_t^n) - \mathrm{E}[u_t(c_t^n)] \\ &= u_t(c_t^n) + \alpha_n V_\pi^n(\mathbf{x}_{t+1}^n) - V_\pi^n(\mathbf{x}_t^n)\end{aligned}$$

である TD 誤差 δ_t を Actor へ伝えると同時に，活性度の履歴 e_{jt}^n

$$e_{jt}^n = \lambda_e e_{j(t-1)}^n + g_j^{C_n}(\mathbf{x}_t^n)$$

を計算し，結合荷重を

$$W_{jt}^{C_n} = W_{j(t-1)}^{C_n} + \eta_C \delta_t e_{jt}^n$$

で更新する．Actor では結合荷重を

$$W_{ijt}^{A_n} = W_{ij(t-1)}^{A_n} + \eta_A \delta_t g_j^{A_n}(\mathbf{x}_t^n) \times rnd_{it}$$

で更新する．ここで，η_A，η_C は学習率，λ_e は履歴の減衰率を表す．

4. シミュレーション結果並びに考察

ここで，人工市場において従来の DP 法によるポートフォリオ戦略の構築が困難と考えられる場合に，ニューラルネットを用いた強化学習を無危険資産が一つ，危険資産が三つ存在する人工市場へ適用した結果を示す．そこで，Actor と Critic へ入力される環境の状態 $\mathbf{x}_t^n$ は $[z_{1t}^n/z_{1(t-1)}^n,\ r_{2t},\ r_{3t},\ r_{4t},\ c_t^n/c_{t-1}^n]^{\mathrm{T}}$ の形で与える．ここで，T はベクトルの転置である．Actor と Critic の動径基底関数はともに各変数を -0.5 から 2.5 の間を 7 等分した値の組合せ（つまり，$N_A = N_C = 7^5$）に中心 $\mathbf{m}_j^{A_n}, \mathbf{m}_j^{C_n}$ を配置した．$\mathbf{C}_j^{A_n}$，$\mathbf{C}_j^{A_n}$ はすべて 7 行 7 列の単位行列とした．取引回数を 2100 回とし，最後の 100 回分を 1 試行として計測した．各エージェントの初期の資産は $1000 + 50\sigma$，各危険資産の初期の価格は $100 + 5\sigma$ で与えた．また，Actor と Critic の初期の結合荷重はともに σ で与えた．ここで，σ は $-1 \sim 1$ の一様乱数である．これらの初期値は試行のたびに新しく乱数を用いて与えている．配当と労働収入は常に 0 であると仮定した．

まず，エージェント数が 9 人，価格の感度はすべての危険資産について $T_i^P = 50,\ (i = 1, 2, 3)$，エージェントの感度はすべて $T_n^A = 1,\ (n = 1, 2, 3, \cdots, 9)$ とし，その他のパラメータを表 1 のように設定した場合の危険資産の価格と期間 t における効用関数 $u_t(c_t)$ の値の変動の一例を図 3，図 4 にそれぞれ示す．ただし，期間 t における効用関数の値は同じ 1 試行において計測した最後の 100 回分の平均が最小，中間，最大となるエージェント 3 人分を表している．

表 1　パラメータの値
Table 1　Values of parameters.

η_A	η_C	λ_e	r_{1t}	α	γ
0.001	0.001	0.8	0.01	0.5	0.5

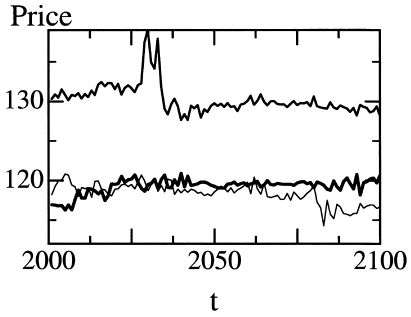

図 3　危険資産の価格の変動
Fig. 3　Change of risk assets.

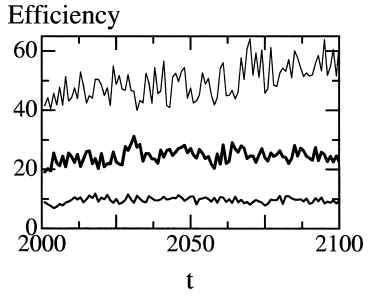

図 4　効用関数の値の変動
Fig. 4　Change of value of utility function.

次に，表 1 の将来の報酬に対する割引率を表す α_n と効用関数の選好の程度を表す γ_n の部分を変えてその影響について調べる．そこで，すべてのエージェントが同じ割引率 α と選好の程度 γ の組合せをもつ場合 $(\alpha,\gamma) = (0.1, 0.1)$, $(0.1, 0.9)$, $(0.5, 0.5)$, $(0.9, 0.1)$, $(0.9, 0.9)$ について，得られた危険資産の価格と効用関数の値を図 5，図 6 に示す．ここでは，各危険資産と各エージェントごとについて各試行において計測された 100 回分の平均値の 10 試行の平均を求めている．各線分の中心が平均を表し，上下の範囲はそれぞれ標準偏差の 3 倍を示す．標準偏差は各危険資産と各エージェントごとについて各試行において計測された 100 回分の標準偏差の 10 試行の平均で求めている．図 5，図 6 の結果から，α が危険資産の価格や効用関数の値に大きな影響を与えないのに対して，γ は大きくなると危険資産の価格と効用関数の値が増加し，併せて効用関数の分散が大きくなることが分かる．

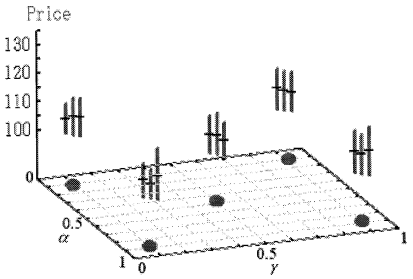

図 5　すべてのエージェントが同じ α や γ の値をもつ場合の価格
Fig. 5　Prices when all agents take the same value about each α and γ.

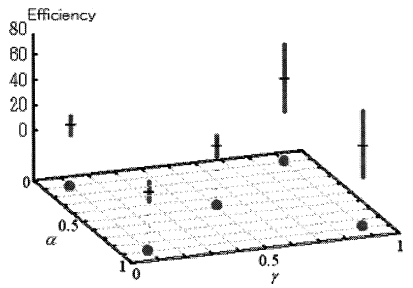

図 6　すべてのエージェントが同じ α や γ の値をもつ場合の効用
Fig. 6　Utilities when all agents take the same value about each α and γ.

表 2　α と γ の組合せ
Table 2　Set of α and γ.

	ケース 1		ケース 2		ケース 3	
	α	γ	α	γ	α	γ
1	0.1	0.1	0.1	0.9	0.4	0.6
2	0.2	0.2	0.2	0.8	0.7	0.5
3	0.3	0.3	0.3	0.7	0.2	0.9
4	0.4	0.4	0.4	0.6	0.8	0.7
5	0.5	0.5	0.5	0.5	0.5	0.1
6	0.6	0.6	0.6	0.4	0.1	0.4
7	0.7	0.7	0.7	0.3	0.6	0.8
8	0.8	0.8	0.8	0.2	0.3	0.2
9	0.9	0.9	0.9	0.1	0.9	0.6

更に，エージェントが異なる割引率 α と選好の程度 γ の組合せをもつ場合について考えた（表 2 参照）．ここで，ケース 1 は最初のエージェントから最後のエージェントにかけて α と γ がともに増大する組合せ，ケース 2 は最初のエージェントから最後のエージェントにかけて α は増大するが γ は減少する組合せ，ケース 3 はランダムな組合せとする．エージェントごとに得られた危険資産の価格をすべてのケースについてま

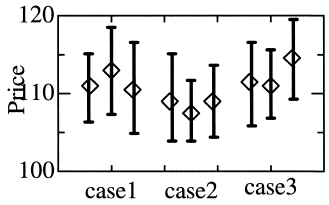

図 7　危険資産の価格
Fig. 7　Prices of risk assets.

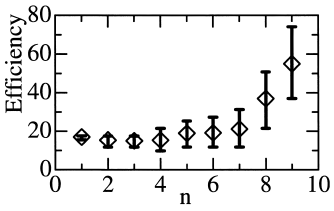

図 8　エージェントの効用（ケース 1）
Fig. 8　Utilities of agents. (Case 1)

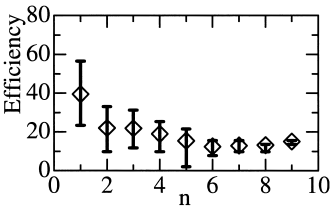

図 9　エージェントの効用（ケース 2）
Fig. 9　Utilities of agents. (Case 2)

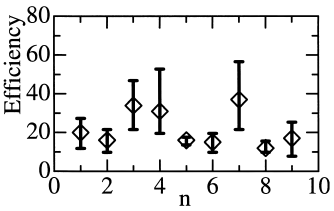

図 10　エージェントの効用（ケース 3）
Fig. 10　Utilities of agents. (Case 3)

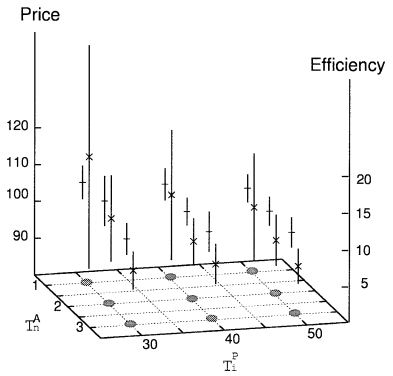

図 11　エージェント数と感度の影響 ($N = 10$)
Fig. 11　Influence of agent number and sensitivity. ($N = 10$)

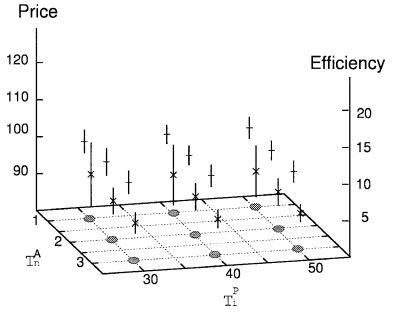

図 12　エージェント数と感度の影響 ($N = 20$)
Fig. 12　Influence of agent number and sensitivity. ($N = 20$)

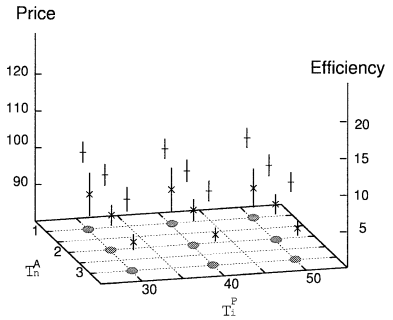

図 13　エージェント数と感度の影響 ($N = 30$)
Fig. 13　Influence of agent number and sensitivity. ($N = 30$)

とめたものを図 7 に示し，ケースごとの効用関数の値を図 8，図 9，図 10 にそれぞれ示す．

これらの結果から，危険資産の価格はそれぞれのケースにおいて大きく異なることがないのにもかかわらず，効用関数の選好の程度を表す γ の値が大きいエージェントほど高い効用関数の値が得られていることが分かる．また，そのばらつきも大きなものとなることも示される．

最後に，危険資産に関する価格の感度 T_i^P やエージェントの感度 T_n^A と，取引に参加するエージェントの数を変更した場合に得られた結果について述べる．その他のパラメータは表 1 で与える．まず，感度については T_i^P が 30, 40, 50，T_n^A が 1, 2, 3 となる 9 通りの組合せを考えた．それぞれの感度の組合せについ

て，エージェント数が 10, 20, 30 となる場合についてシミュレーションを行った結果，図 11，図 12，図 13 をそれぞれ得た．図中において，各 $(T_i^P,\ T_n^A)$ の組合せに対する危険資産の価格の振舞いを左上側に，全エージェントの効用関数の振舞いを右上側に組にして示している．ここでは，まず各危険資産と各エージェントごとについて各試行において計測された 100 回分の平均値の 10 試行の平均を求めた上で，すべての危険資産やすべてのエージェントの平均を求めている．各線分の中心が平均を表し，上下の範囲はそれぞれ標準偏差の 3 倍を示す．標準偏差は各危険資産と各エージェントごとについて各試行において計測された 100 回分の標準偏差の 10 試行を求めた上で，すべての危険資産やすべてのエージェントの平均で求めている．

これらの結果から，危険資産の数に対して市場に参加するエージェント数が増加すると危険資産の価格や効用関数の値のばらつきが抑制される傾向があることが分かる．このことは，エージェント 1 人が価格形成に及ぼす影響が減少するためと考えられる．また，危険資産に関する価格の感度 T_i^P が大きくなると，危険資産の価格の値がわずかながら大きくなることが分かる．更に，エージェントの感度 T_n^A が小さくなると，エージェントの効用関数の値のばらつきが大きくなることが分かる．このことは，感度 T_n^A が小さくなると Actor と Critic への入力の一部である消費の変化に敏感になるためであると考えられる．

5. むすび

本研究では，投資家の合理的な動作を強化学習をエージェントに適用することで仮想的にコンピュータ上で実現した．更に，エージェントのもつ将来の報酬に対する割引や効用関数に対する選好の程度が形成される金融価格や消費に関する期待効用へ及ぼす影響について分析した．提案手法を洗練していくことで，強化学習を利用したポートフォリオ支援システムの開発に役立てることが考えられる．本研究では限られた状況に対して得られた結果を示しているが，その他のパラメータが及ぼす影響について分析することが今後の課題といえる．

文　献

[1] 沢木勝茂，ファイナンスの数理，朝倉書店，1995.

[2] F. Black and M. Scholes, “The pricing of options and corporate liabilities,” J. Political Economy, vol.81, pp.637–654, 1973.

[3] R.G. Palmer, W.B. Arthur, J.H. Holland, B. LeBaron, and P. Tayler, “Artifical economic life: A simple model of a stockmarket,” Physica D, vol.75, pp.264–274, 1994.

[4] R.S. Sutton and A. Barto, “Reinforcement learning: An introduction,” A Bradford Book, The MIT Press, 1998.

[5] 今野　浩，古川浩一，ファイナンスハンドブック，朝倉書店，1997.

[6] N. Hakansson, “Optimal investiment and consumption strategies under risk for a class of utility function,” Econometrica, vol.38, pp.587–607, 1970.

[7] R. Bellman, Dynamic programming, Princeton University Press, Princeton, 1957.

[8] A.G. Barto, R.S. Sutton, and C.W. Anderson, “Neuronlike adaptive elements that can solve difficult learning control problems,” IEEE Trans. Syst. Man. Cybern., vol.13, no.5, pp.835–846, 1983.

[9] K. Doya, “Reinforcement learning in continuous time and space,” Neural Computation, vol.12, pp.243–269, 1999.

（平成 15 年 7 月 11 日受付，10 月 7 日再受付，
16 年 1 月 23 日最終原稿受付）

奥原　浩之　（正員）

平 8，広島大大学院工学研究科システム工学専攻博士課程後期了．同年広島大学工学部助手．平 10 広島県立大学経営学部講師，平 12 助教授．博士（工学）．神経回路網工学とその応用研究に従事．システム制御情報学会，日本 OR 学会，日本ファジイ学会各会員．

柴田　淳子　（学生員）

平 14 広島県立大学経営情報学研究科修士課程了．同年，広島大学工学研究科複雑システム工学専攻博士課程後期入学．主としてマルチエージェントの研究に従事．

田中稔次朗　（正員）

昭 51 阪大・応用物理博士課程了．昭 52 鹿児島県立短期大講師．昭 53 同助教授．昭 59 同教授．平 9 広島県立大・経営教授．工博．カオス理論とその工学への応用研究に従事．昭 54〜55 カリフォルニア大学サンタバーバラ校客員研究員．平 2MBC 賞受賞．日本物理学会，アメリカ物理学会，情報処理学会各会員．

坂和　正敏　（正員）

昭 45 京大・工・数理卒．昭 50 同大大学院博士課程了．昭 50 神戸大・工・システム助手．昭 56 同助教授．昭 62 岩手大・工・教授．平 2 広島大・工・第二類（電気系）計数管理工学講座教授を経て，平 13 同大大学院工学研究科教授となり現在に至る．大規模システム，多目的システム，ファジーシステムにおける意思決定法手法とその応用に関する研究に従事．工博．著書「線形システムの最適化」,「非線形システムの最適化」,「ファジィ理論の基礎と応用」,「経営数理システムの基礎」（森北出版）,「ソフト最適化」,「遺伝的アルゴリズム」（朝倉書店）, Fuzzy Sets and Interactive Multiobjective Optimization（Plenum Press）等．