強化学習を用いたライフサイクル投資

Life-Cycle Investment with Reinforcement Learning

上田 翼 [1*]
Tsubasa Ueda[1]

[1] 三井住友 DS アセットマネジメント株式会社
[1] Sumitomo Mitsui DS Asset Management Company, Limited

Abstract: 個人のライフサイクルを前提として長期投資を考える場合，通常の分散分析アプローチとは異なり多期間最適化の視点が必要となる．より現実に即した仮定の下で，強化学習を用いて構築した投資ポートフォリオは典型的な手法より優れた生涯効用を達成し得ることを検証する．

1 はじめに

個人の生涯を通じた資産配分を考えるライフサイクル投資と呼ばれる分野がある．通常の平均分散アプローチでは，投資期間は 1 期のみであり，期末に全ての富を消費することが暗黙裡に仮定されている．しかしながら、現実の消費は複数の期間にわたって行われるため、長期投資においては毎期の消費と投資ポートフォリオを同時に考慮する多期間最適化の視点が必要となる．比較的単純な仮定の下では，Merton[1] が連続時間，Samuelson[2] が離散時間の枠組みにおいて最適消費比率，リスク資産への最適投資比率の解析解を明らかにした．その後，多くの研究によって，効用関数の修正，消費制約の追加，投資資産の拡大といった拡張が試みられてきた．しかし，大半は状態と行動を離散化したうえで価値関数をモンテカルロ法で計算しており，離散化基準の恣意性やリスク情報の欠落，扱える次元の制約といった課題が存在する．

近年発展している深層強化学習の分野では，価値関数をニューラルネットワーク等で近似することで，連続な状態行動空間において幅広い問題を解くことが可能となった．それゆえ，より柔軟な仮定の下でライフサイクル投資にアプローチできる可能性が期待される．実際に [3] では，GJR-GARCH 型の株価と Hull-White モデルに従う金利などを仮定し，強化学習を用いて退職世代の最適消費・投資を分析している．

本稿では，金利のダイナミクスとして実務で一般的な Nelson-Siegel 型のイールドカーブモデルを導入したうえで，労働収入により資産を蓄積する現役世代まで対象を広げ，消費と投資ポートフォリオの分析を行う．強化学習モデルによる投資戦略は，60/40 ポートフォリオやターゲット・デート・ファンドなどの典型的な手法よりも優れた生涯効用を達成し得ることを検証する．

*tsubasa_ueda@smd-am.co.jp, tsubasa.ud@gmail.com

2 ライフサイクルと資産価格モデル

2.1 ライフサイクル

個人は 22 歳で保有資産 0 の状態から労働を開始し．賃金収入を得て消費を行いながら余剰資産を投資する．65 歳で退職して以降は年金を受給しつつ，現役世代と同様に消費と投資を続ける．投資対象は株式と債券（デュレーション 1〜30 年を同時に選択）であり，消費および投資ポートフォリオの意思決定は年に 1 回行われる．最大寿命を 105 歳とし，一般的な死亡率に従い毎期確率的に逝き．残された資産は遺産となる．

2.2 効用関数

消費効用関数として，次のような相対的リスク回避度 γ が一定である関数を仮定する．

$$u_c(c) = \frac{c^{1-\gamma} - 1}{1-\gamma} \quad (c > 0)$$

同様に，遺産効用関数として，次のような相対的リスク回避度 γ が一定で遺産選好 ϕ をもつ関数を仮定する．

$$u_b(b) = \frac{b^{1-\gamma}}{1-\gamma}\left(\frac{\phi}{1-\phi}\right)^{\gamma} - \frac{1}{1-\gamma} \qquad (b > 0)$$

x 歳を開始年齢として，消費効用と遺産効用を合わせた期待生涯効用を次のように定義する．[4]

$$E\left[\sum_{t=0}^{T} {}_tp_x u_c(c_t) + {}_{t-1|}q_x u_b(b_t)\right] \tag{1}$$

${}_tp_x$ は x 歳時に生存している場合に $x+t$ 歳に生存している確率であり，${}_{t-1|}q_x$ は x 歳時に生存している場合に $(x-t-1, x-t]$ 歳で死亡する確率である．個人は期待生涯効用を最大化するように意思決定を行う．なお，本稿では $\gamma = 10$ とあらかじめ設定している．

2.3 収入

現役時の賃金は一般的な大学学卒男性の賃金カーブを想定し，退職時には初任給対比で平均 8 倍の退職金を受け取ると仮定した．毎年の賃金は景気による賞与変動を捉えるため株価と相関 $\rho_{stock,wage} = 0.3$ を持つ誤差項 ϵ_{wage} を加え，退職金には個人の異質性を反映して独立な誤差項 $\epsilon_{package}$ を加える．退職後は初任給対比で 75%の固定的な年金を受給し続ける．

2.4 株式

株式は，年金積立金管理運用独立行政法人（GPIF）の基本ポートフォリオが依拠する前提 [5] を参考に，年間期待リターンを 5.6 %，年間標準偏差を 23.14 %とした. リターン分布は正規分布を仮定した．

2.5 債券

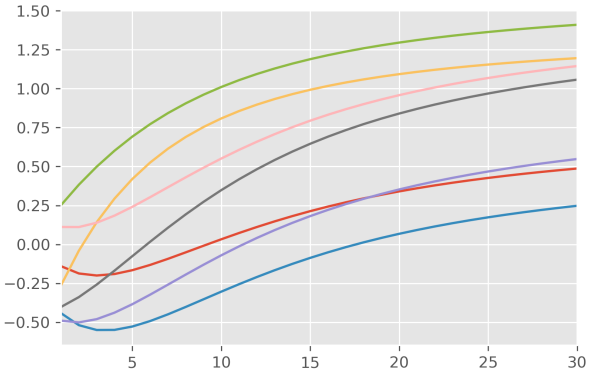
イールドカーブのダイナミクスとして，次のような動学的 Nelson-Siegel モデルを想定する．[6]

$$r_{\tau,t} = \beta_{t,1} + \beta_{t,2}\frac{1-e^{\lambda\tau}}{\lambda\tau} + \beta_{t,3}(\frac{1-e^{\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}) + \epsilon_{\tau,t}$$

$$\boldsymbol{\beta_t} - \boldsymbol{\mu} = \boldsymbol{A}(\boldsymbol{\beta_{t-1}} - \boldsymbol{\mu}) + \boldsymbol{\xi_t}$$

$r_{\tau,t}$ は時点 t の年限 τ の金利を指し，$\boldsymbol{\beta_t} = (\beta_{t,1}, \beta_{t,2}, \beta_{t,3})$ は $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ に平均回帰する VAR(1) 過程に従う. 誤差項 $\epsilon_{\tau,t}$ は互いに独立だが，$\xi_{t,j}(j = 1, 2, 3)$ は互いに相関を仮定する．状態空間モデルを用いて，日本の 2000 年 1 月～2020 年 8 月の月次データをもとにパラメーターを推定した．生成したイールドカーブの 7 年間の推移例を図 1 に示す．債券はイールドカーブの変化に従い価格が変動するためリスク性資産となるが，デュレーションが短期化するにつれてリスクは低下していく. 特に 1 年債は投資時点でリターンが確定することから実質的に無リスク資産とみなせる．

図 1: イールドカーブの推移例



3 強化学習モデル

3.1 状態行動空間

状態空間と行動空間を次のように定義する．

$$\begin{aligned} s_t \in \mathcal{S} &:= \mathcal{Z} \times \mathcal{R} \times \mathcal{R}^3 \\ &= \{(l, w, \boldsymbol{\beta}) \mid l, w \geq 0\} \\ a_t^{\pi}(s_t) \in \mathcal{A} &:= \mathcal{R} \times \mathcal{R} \times \mathcal{R} \\ &= \{(\alpha, \delta, \tau) \mid 0<\alpha<1, 0\leq\delta\leq2, 1\leq\tau\leq30\} \end{aligned}$$

l は最大余命，w は保有資産，$\boldsymbol{\beta}$ は前述したイールドカーブの状態を表している．エージェントは毎期これらの状態を観察し，以下の行動を決定する．α は資産に占める消費の比率である．δ は投資に占める株式の比率であり，最近のレバレッジ型投信の普及を踏まえて，δ の上限は 2 としている．ただし，レバレッジ部分については手数料率 $r_{leverage} = 0.01$ を支払う．τ は債券の投資デュレーションであり，最大を 30（年）とした．

3.2 アルゴリズム

学習には PPO(proximal policy optimization) を用いる．生涯期待効用 (1) の最大化に対応するよう，開始時点の年齢 $x = 22$ として報酬を次のように定義する．

$$r_t = {}_tp_x u_c(c_t) + {}_{t-1|}q_x u_b(b_t)$$

PPO の方策関数と価値関数は，ユニット数 64，レイヤー数 2 の隠れ層からなる全結合型ニューラルネットワークで構成する．PPO のパラメータは次の目的関数を最大化するように更新される．[7]

$$L_t^{CLIP+VF+S}(\theta) = \hat{E}_t[L_t^{CLIP}(\theta) - c_1 L_t^{VF}(\theta) + c_2 S[\pi_\theta](s_t)]$$

L_t^{CLIP} はクリップされた代理目的関数，L_t^{VF} は価値関数の二乗誤差，S はエントロピーボーナスである．c_1 は価値関数係数，c_2 はエントロピー係数であり，ハイパーパラメータとして扱われる．

4 シミュレーション

4.1 手順

前述のライフサイクルと資産価格モデルに基づいて，学習用データを 300,000 エピソードを生成する．各エピソードは開始年齢 22 歳から最大寿命 105 歳に至る 84 ステップから構成される．前述の強化学習モデルを用いて学習を行う．学習完了後，テストデータとして 1,000 エピソードを生成する．学習済みのモデルに加え

て，ベンチマークとして①60/40 ポートフォリオと②シンプルなターゲット・デート・ファンドのケースも評価を行う．

4.2 学習結果

図 2 と図 3 は，それぞれ資産と消費の生涯パスの 5%点，中央値，95%点を表している．横軸は労働開始からの経過年に対応する．なお，相対的にリスク回避度一定の効用関数を仮定しているため，縦軸の水準はスケール不変である．資産は，労働を開始し 10 年程度経過してから徐々に蓄積が始まり，退職給付を得た時点が概ねピークとなって次第に取り崩される．消費は退職時まで概ね横ばいで推移し，退職後の資産状況に応じて徐々に増やしていく形となっている．

図 2: 資産

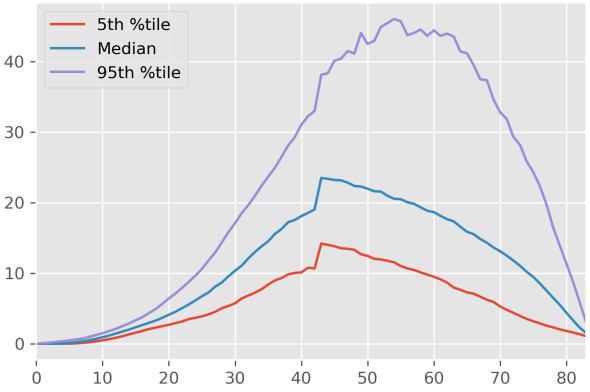


図 3: 消費

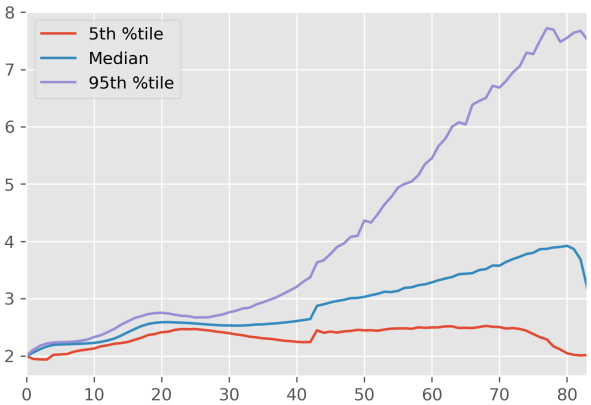


図 4 と図 5 からは投資ポートフォリオの状況がわかる．株式投資比率は労働開始後 10 年から 20 年の間に急上昇し，以降は横ばいで推移する．晩年に至ると比率が急低下しているが，資産が取り崩されてバッファーが減ったことや遺産動機が相対的に優勢になることが関係していよう．債券デュレーションは，当初は高水準ながら退職時にかけて緩やかに低下しており，ポートフォリオ全体のリスクを抑制している．株式投資比率ではなくデュレーションを用いたリスクコントロールはあまり一般的ではないが，本稿では平均回帰型のイールドカーブモデルを用いていることから債券価格にわずかな予測可能性が生まれるため，安定的な投資として債券が選好された可能性があろう．晩年になるとデュレーションは再び上昇するが，株式投資比率の低下と対になっており同じ背景があると考えられる．

図 4: 株式投資比率

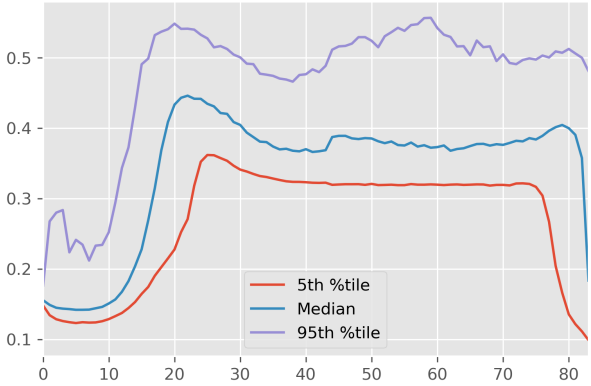
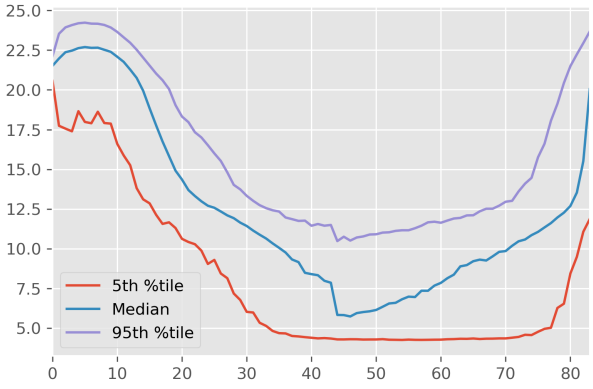


図 5: 債券デュレーション



4.3 ベンチマーク比較

学習した消費モデルを固定しつつ，投資モデルのみベンチマークに変更して評価を行う．60/40 ポートフォリオは主に米国で退職世代の安定的なポートフォリオとして推奨されており，株式投資比率 δ は 0.6 に固定する．ターゲットデート型ファンド（TDF）は年齢が上がるにつれて株式投資比率を低下させていくが，今回は”rule of thumb”として有名な $\delta = 1 - $年齢$/100$ を用いる．ベンチマークの債券デュレーション τ は，ラ

ンダムに選択するものとする．

生涯効用の平均値と標準偏差は表 1 の通りである．学習モデルの平均が最も高く，標準偏差が最も低い．60/40 ポートフォリオは平均が最も低く，標準偏差が最も高い．TDF の平均は学習モデルと遜色ないが．標準偏差がやや高く，資産価格の変動に左右されやすい．以上のことから，強化学習により構築した投資戦略は，典型的な投資ポートフォリオより生涯効用の観点で優位になり得るといえよう．もっとも，これらの投資戦略の違いがもたらす生涯効用の差自体はそれほど大きくなく，生涯効用は消費（貯蓄）のあり方の影響が大きい可能性もあろう．

表 1: 生涯効用

	学習モデル	60/40	TDF
平均	7.41275	7.41251	7.41274
標準偏差	0.00066	0.00152	0.00073

5 結論

本稿では，強化学習を用いて，ライフサイクル投資の観点から最適な消費とポートフォリオを検討した．より現実に即した，柔軟性の高い仮定と連続な状態行動空間のもとで，学習したモデルは典型的なポートフォリオと比べて優れた生涯効用を獲得した．もっとも，資産価格のダイナミクスや学習アルゴリズムを網羅的に検討したわけではなく，結果は幅をもってみる必要がある．

参考文献

[1] Merton, Robert C. ”Lifetime portfolio selection under uncertainty: The continuous-time case.” *The review of Economics and Statistics* (1969): 247-257.

[2] Samuelson, Paul A. ”Lifetime portfolio selection by dynamic stochastic programming.” *The review of economics and statistics* (1969): 239-246.

[3] Gordon Irlam. ”Machine Learning for Retirement Planning.” *The Journal of Retirement* 8 (1) (2020): 32-39.

[4] Bell, David, Estelle Liu, and Adam Shao. ”Member’s Default Utility Function for Default Fund Design Version 1 (“MDUF v1”) Technical Paper No. 3: Optimal Dynamic Strategies.” dynamics 50 (2017): 2-2.

[5] 年金積立金管理運用独立行政法人．”基本ポートフォリオの変更について（詳細）” https://www.gpif.go.jp/topics/Adoption%20of%20New%20Policy%20Portfolio_Jp_details.pdf.

[6] Diebold, Francis X., and Canlin Li. ”Forecasting the term structure of government bond yields.” Journal of econometrics 130.2 (2006): 337-364.

[7] Schulman, John, et al. ”Proximal policy optimization algorithms.” arXiv preprint arXiv:1707.06347 (2017).

[8] 安達智彦, 池田昌幸: 『長期投資の理論と実践』東京大学出版会 (2019)

[9] Gourinchas, Pierre- Olivier, and Jonathan A. Parker. ”Consumption over the life cycle.” Econometrica 70.1 (2002): 47-89.