強化学習を用いた期待効用ベースヘッジ手法

Expected Utility Based Hedge 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: Selling options is a popular investment strategy, which regularly receives a premium and, on the other hand, takes variance risk, especially negative fat-tail risk. Therefore, it is important for risk-averse investors to mitigate these types of risks by constructing hedge position in consideration of transaction costs. Main results of this research are as follows: (1) In a practical simulation, DDPG model with utility based reward suggests a better way of dynamic hedging compared to simple benchmarks. (2) As a real-world application to market data, this learned model successfully manages the short straddle portfolio of treasury futures options.

1 はじめに

投資戦略のパフォーマンスを測るうえでは、期待リターンとボラティリティを基準に考えるのが一般的である．ある種の仮定の下では、平均分散分析によってポートフォリオを最適化すると、投資家の期待効用が最大化されるためである．しかしながら、投資家の効用関数が二次以外、かつ投資収益が正規分布しないときには、平均分散分析が必ずしも効率的にならないことが知られている [1]．前者の二次効用関数は単純化された仮定であり、現実的な投資家の行動を考えると拡張の余地は大きい．後者のように収益分布が非正規となる典型例はオプションやボラティリティ・インデックス（VIX 等）のポジションであり、市場が急変すれば損益が非線形に拡大しうる．このような戦略を適切に評価するには、平均分散分析では不十分であり、収益分布の歪度等を考慮した期待効用で考える必要がある．

一方、近年の技術的な動向においては、オプション・リスクの最適ヘッジ量を推定する手段として、深層学習が注目されている．[2] では、Semi-RNN を用いて凸リスク測度に基づくヘッジが提案された。[3] は、非線形の取引コストを仮定したうえで、DQN によるヘッジを扱っている。他方、著者の印象では、モデルの目的関数は実務的な観点から先験的に定義されたものが多いほか、実証面でも数値実験にとどまり現実の市場や投資戦略を必ずしも反映していないケースが多い。

そこで本稿では、期待効用理論の観点から自然な形で報酬を設計したうえで深層強化学習モデルを構築し、実際の市場データと投資戦略を用いて学習の妥当性を検証する．

*E-mail: tsubasa.ud@gmail.com

2 シミュレーション

2.1 期待効用モデル

投資家は、次のような絶対的リスク回避度 γ の指数型効用関数をもつと仮定する．

$$u(x) = \frac{1}{\gamma}(1 - e^{-\gamma x}) \tag{1}$$

x の平均 μ の周りでテイラー展開を行ったうえで期待値を取ると、

$$\begin{aligned} E[u(x)] &= u(\mu) - \frac{1}{2}u''(\mu)E[(x-\mu)^2] \\ &\quad + \frac{1}{6}u'''(\mu)E[(x-\mu)^3] + \dots \\ &\approx \mu - \frac{\gamma}{2}\mu_2[x] + \frac{\gamma^2}{6}\mu_3[x] \end{aligned} \tag{2}$$

と書ける。ここで、(2) は $\mu = 0$ 周りの近似であり、$u(\mu) \approx u(0) + u'(0)\mu = \mu$ を用いた．なお、μ_n は n 次中心モーメントである。ある T 時点の富を w_T とし、(1) を参照点 w_0 に依存する効用関数と捉えると、期待効用 $E[u(x)]$ の最大化は

$$\max\Big[E[w_T - w_0] - \frac{\gamma}{2}\mu_2[w_T - w_0] + \frac{\gamma^2}{6}\mu_3[w_T - w_0]\Big] \tag{3}$$

と表せる．$w_T - w_0 = \sum_{t=1}^{T} \Delta w_t$ と分解し、$Cov(\Delta w_t, \Delta w_s) = 0\,(t \neq s)$ を仮定すると、モーメントの加法性から、(3) は

$$\max \sum_{t=1}^{T}\Big[E[\Delta w_t] - \frac{\gamma}{2}\mu_2[\Delta w_t] + \frac{\gamma^2}{6}\mu_3[\Delta w_t]\Big] \tag{4}$$

と変形できる．[4] の議論を参考にすると、$\mu = E[\Delta w_t]$ が相対的に小さいならば $\mu_2 = E[(\Delta w_t)^2] - \mu^2$, $\mu_3 = E[(\Delta w_t)^3] - 3\mu\mu_2 - \mu^3$ の μ^2, μ^3 を無視できて、(4) は次のような報酬を設定した強化学習の問題として近似できる．

$$r_t = \mu - \frac{\gamma}{2}(\Delta w_t)^2 + \frac{\gamma^2}{6}\{(\Delta w_t)^3 - 3\mu(\Delta w_t)^2\} \quad (5)$$

ただし、μ は事前には明らかでないので後に置き換える．

2.2 環境とモデル

$t = 0$ 時点でオプションのポジションを構築し、離散時間 t ごとに原資産でヘッジする戦略を想定する．原資産価格が Heston モデル [5] に従うとして、状態空間と行動空間を次のように定義する．

$$\begin{aligned} s_t \in \mathcal{S} &:= \mathcal{R}_{++}^2 \times \mathcal{R}_+ \times \mathcal{R} \\ &= \{(S, \nu, \tau, \delta) \mid S, \tau > 0, \delta \geq 0, -1 \leq \delta \leq 1\} \\ a_t = a_t^\pi(s_t) &\in \mathcal{A} := \mathcal{R} \end{aligned}$$

S, ν, τ, δ はそれぞれ、原資産価格、原資産価格のインプライドボラティリティ、オプションの残存期間、ヘッジ比率を表す．a は定常な決定的方策 π に従って選択されたヘッジ資産の売買量であり、$\delta_{t+1} = \delta_t + a_t$ という関係が成立する．a_t から生じる取引コストを c_t、ヘッジなしオプション戦略の期待リターンを μ_{un} と表し、ヘッジ資産の期待収益を中立とみなせば、$\mu \approx \mu_{un} - E[c_t]$ と近似できるから、期待効用ベースの報酬 (5) を

$$\begin{aligned} r_t = &- c_t - \frac{\gamma}{2}(\Delta w_t)^2 \\ &+ \frac{\gamma^2}{6}\{(\Delta w_t)^3 - 3(\mu_{un} - c_t)(\Delta w_t)^2\} \end{aligned} \quad (6)$$

と置き換える．なお、$Cov(c_t, (\Delta w_t)^2) = 0$ を仮定し、μ_{un} を $a_t = 0$ $(0 \leq t \leq T)$ とおいて事前に推定しておくものとする．以上の環境は、連続状態行動空間のマルコフ決定過程 $M(\pi)$ となる．

目的関数として次の期待報酬和を設定し、DDPG(Deep Deterministic Policy Gradient) 法 [6] を適用する．

$$E\Big[\sum_{t=1}^{T} r_t \mid M(\pi)\Big] \quad (7)$$

DDPG の Actor および Critic は深層ニューラルネットワークで構成し、方策勾配法によりパラメータを更新する．

2.3 データセットと戦略

投資対象として米国債券先物オプション、ヘッジ資産として米国債券先物を想定する．離散化した Heston モデルにより、原資産価格とオプション価格のサンプルデータを、学習用に 400,000 ステップ、テスト用に 400,000 ステップ生成した．次章の実証分析を念頭において、Heston モデルにはリスクプレミアムの要素を導入し [7]、パラメータは実際の市場に近いものを設定した．

1 エピソードは、$S_0 = 100 + \epsilon$ と微小なランダム性を加えて基準化したうえで、$t = 0$ 時点で行使価格 $K = 100$ のショートストラドル（同一行使価格でのプットとコールの売りポジション）を構築し、t ごとにヘッジ比率 δ_t を調整しつつ、満期 $T = 40$ まで保有して終了する．オプションの満期前行使は想定せず、満期時に ITM(In The Money) であれば原資産に転換されるが、同時にポジション解消に必要な売買を行うものとする．取引コスト c_t はヘッジ資産の売買量 a_t に比例すると仮定した．保有するオプションとヘッジ資産の価格変化を反映したポートフォリオ収益を Δw_t とする．$R = w_T - w_0$ とすれば、期待報酬和 (7) の最大化は、効用関数 (1) に基づく収益の期待効用 $E[u(R)]$ を近似的に最大化していると解釈できる．

2.4 学習結果

図 1 は、テストデータで 10,000 回シミュレーションした収益 R の分布である．比較対象として、ヘッジなし (UN) と Black モデルに基づくデルタ（オプションの価格変化／原資産の価格変化）を毎時点で中立化する方法 (BD) を用意した．UN では、当然ながらボラティリティが大きく負のファットテールが目立っている．BD では、ボラティリティを大幅に抑制できているものの、取引コストを考慮せずヘッジ資産を頻繁に売買するため収益分布の中心が 0 に近づいている．強化学習モデルに基づくヘッジ（RL）では、収益分布の中心をある程度プラスに維持したままボラティリティとファットテールを適度に抑制できている．

同じことを統計量で確認したのが、表 1 である．平均と分散は t を日次とみなして R/S_0 を年率換算し、平均効用 $\overline{u(R)}$ は (2) の近似式ベースで計算した．BD と比べて RL の収益は分散や歪度が多少劣後するものの、平均収益率が高いため平均効用で僅かに上回る結果となった．

図 (2)、(3) は $\delta_{t-1} = 0$ とした際の RL と BD のヘッジ資産売買量を比較したものである．$0 \leq \tau(t) \leq 0.1$, $-2 \leq (S_t - K) \leq 2$ の範囲における a_t（垂直軸）を表している．RL は BD より a_t の絶対値が小さく、

コストを考慮してヘッジ資産の調整幅を制限していることがわかる．また、BD の a_t は正負が対称的になっているが、RL の a_t は正側が相対的に大きくなっており、原資産の下落リスクより上昇リスクを重視していると解釈できる．

表 1: 収益率の統計量（シミュレーション）

	UN	BD	RL
平均（年換算、%）	0.94	0.65	0.87
分散（年換算、%）	2.80	0.30	0.57
歪度	-1.01	0.04	-0.25
平均効用	-0.167	0.047	0.048

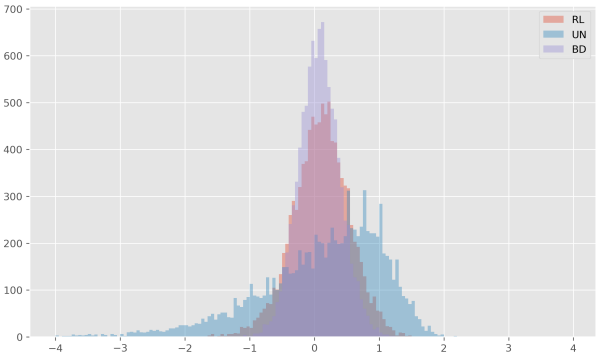

図 1: 収益分布（シミュレーション）

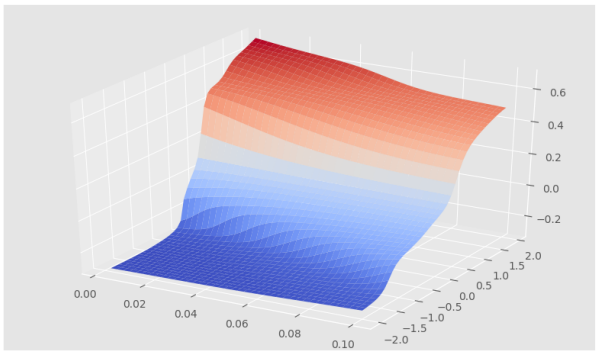

図 2: RL によるヘッジ調整量（シミュレーション）

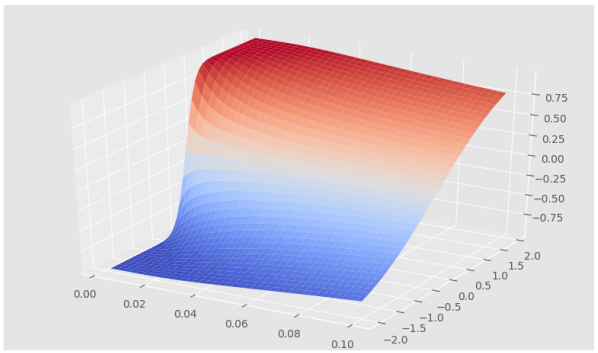

図 3: BD によるヘッジ調整量（シミュレーション）

3 実証分析

3.1 取引ルール

前章で学習した強化学習モデルを実際の市場データに適用しバックテストを行った．サンプル期間は 12 年 5 月～19 年 12 月とし、シミュレーションと同様に期近オプションを売り建てて日次でヘッジしつつ満期まで保有する戦略を繰り返した．なお、収益からはオプションやヘッジに用いる先物の売買コスト・手数料等を控除した．

3.2 バックテスト結果

図 4 は各ヘッジ手法を採用した際の NAV の推移である．UN と比べて RL のボラティリティやドローダウンは比較的抑制されており、長期的な累積リターンもプラスであることから、投資戦略として機能しているといえる．18 年後半からリターンがマイナス化しているが、実際に米金利のボラティリティが高止まりしていたため自然な結果である．

表 2 は、各ヘッジ手法の統計量である．前章のシミュレーション結果と同様に RL が最も高い平均効用を獲得した．なお、BD では平均収益率がほぼ 0 まで低下しており、現実的なヘッジ戦略を考えるうえで取引コストの重要性を示唆している．

表 2: 収益率の統計量（バックテスト）

	UN	BD	RL
平均（年率、%）	1.00	0.05	0.46
分散（年率、%）	2.79	0.72	0.93
歪度	-1.25	0.07	-0.44
平均効用	-0.106	-0.007	0.020

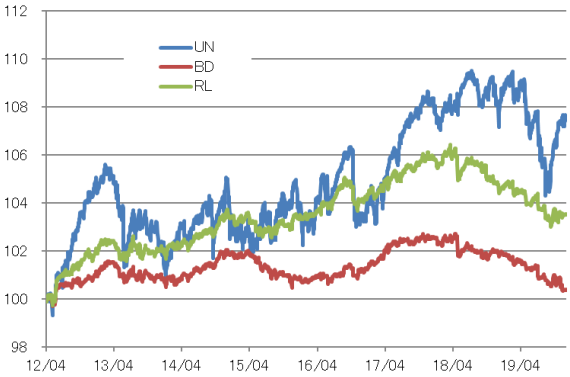

図 4: NAV の推移（バックテスト）

参考文献

[1] 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)

[2] Buehler, Hans, et al. "Deep hedging." *Quantitative Finance* 19.8 (2019): 1271-1291.

[3] Kolm, Petter N., and Gordon Ritter. "Dynamic replication and hedging: A reinforcement learning approach." *The Journal of Financial Data Science* 1.1 (2019): 159-171.

[4] Ritter, Gordon, Machine Learning for Trading (August 8, 2017). Available at SSRN: https://ssrn.com/abstract=3015609 or http://dx.doi.org/10.2139/ssrn.3015609

[5] Heston, Steven L. "A closed-form solution for options with stochastic volatility with applications to bond and currency options." *The review of financial studies* 6.2 (1993): 327-343.

[6] Lillicrap, Timothy P., et al. "Continuous control with deep reinforcement learning." *arXiv preprint arXiv:1509.02971* (2015).

[7] Bollerslev, Tim, Michael Gibson, and Hao Zhou. "Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities." *Journal of econometrics* 160.1 (2011): 235-245.