Annals of Actuarial Science (2023), **17**, pp. 503–546
doi:10.1017/S1748499523000027

ORIGINAL RESEARCH PAPER

Pseudo-model-free hedging for variable annuities via deep reinforcement learning†

Wing Fung Chong[1], Haoen Cui[2] and Yuxuan Li[3]*

[1]Maxwell Institute for Mathematical Sciences and Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh EH14 4AS, UK; [2]School of Computer Science, Georgia Institute of Technology, Atlanta, GA 30332, USA; and [3]Department of Mathematics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
*Corresponding author. E-mail: yuxuanl9@illinois.edu

(Received 08 March 2022; revised 23 January 2023; accepted 06 February 2023; first published online 14 March 2023)

Abstract

This paper proposes a two-phase deep reinforcement learning approach, for hedging variable annuity contracts with both GMMB and GMDB riders, which can address model miscalibration in Black-Scholes financial and constant force of mortality actuarial market environments. In the training phase, an infant reinforcement learning agent interacts with a pre-designed training environment, collects sequential anchor-hedging reward signals, and gradually learns how to hedge the contracts. As expected, after a sufficient number of training steps, the trained reinforcement learning agent hedges, in the training environment, equally well as the correct Delta while outperforms misspecified Deltas. In the online learning phase, the trained reinforcement learning agent interacts with the market environment in real time, collects single terminal reward signals, and self-revises its hedging strategy. The hedging performance of the further trained reinforcement learning agent is demonstrated via an illustrative example on a rolling basis to reveal the self-revision capability on the hedging strategy by online learning.

Keywords: Two-phase deep reinforcement learning; Variable annuities hedging; Training phase; Sequential anchor-hedging reward signals; Online learning phase; Single terminal reward signals; Hedging strategy self-revision.

1. Introduction

Variable annuities are long-term life products, in which policyholders participate in financial investments for profit sharing with insurers. Various guarantees are embedded in these contracts, such as guaranteed minimum maturity benefit (GMMB), guaranteed minimum death benefit (GMDB), guaranteed minimum accumulation benefit (GMAB), guaranteed minimum income benefit (GMIB), and guaranteed minimum withdrawal benefit (GMWB). According to the Insurance Information Institute in 2020, the sales of variable annuity contracts in the United States have amounted to, on average, 100.7 billion annually, from 2016 to 2020.

†This work was first initiated by the authors at the Illinois Risk Lab in January 2020. This work was presented at the 2020 Actuarial Research Conference in August 2020, the United As One: 24th International Congress on Insurance: Mathematics and Economics in July 2021, the 2021 Actuarial Research Conference in August 2021, Heriot-Watt University in November 2021, University of Amsterdam in June 2022, and the 2022 Insurance Data Science Conference in June 2022. The authors thank the participants for fruitful comments. This work utilizes resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign. The authors are grateful to anonymous reviewers for their careful reading and insightful comments. The programming code is publicly available at the GitHub with the following link: https://github.com/yuxuanli-lyx/gmmb_gmdb_rl_hedging

Due to their popularity in the market and their dual-risk bearing nature, valuation and risk management of variable annuities have been substantially studied in the literature. By the risk-neutral option pricing approach, to name a few, (Milevsky & Posner 2001) studied the valuation of the GMDB rider; valuation and hedging of the GMMB rider under the Black-Scholes (BS) financial market model were covered in Hardy (2003); the GMWB rider was extensively investigated by Milevsky & Salisbury (2006), Dai *et al.* (2008), and Chen *et al.* (2008); valuation and hedging of the GMMB rider were studied in Cui *et al.* (2017) under the Heston financial market model; valuation of the GMMB rider, together with the feature that a contract can be surrendered before its maturity, was examined by Jeon & Kwak (2018), in which optimal surrender strategies were also provided. For a comprehensive review of this approach, see Feng (2018).

Valuation and risk management of variable annuities have recently been advanced via various approaches as well. Trottier *et al.* (2018) studied the hedging of variable annuities in the presence of basis risk based on a local optimisation method. Chong (2019) revisited the pricing and hedging problem of equity-linked life insurance contracts utilising the so-called principle of equivalent forward preferences. Feng & Yi (2019) compared the dynamic hedging approach to the stochastic reserving approach for the risk management of variable annuities. Moenig (2021a) investigated the valuation and hedging problem of a portfolio of variable annuities via a dynamic programming method. Moenig (2021b) explored the impact of market incompleteness on the policyholder's behaviour. Wang & Zou (2021) solved the optimal fee structure for the GMDB and GMMB riders. Dang *et al.* (2020) and (2022) proposed and analysed efficient simulation methods for measuring the risk of variable annuities.

Recently, state-of-the-art machine learning methods have been deployed to revisit the valuation and hedging problems of variable annuities at a portfolio level. Gan (2013) proposed a three-step technique, by (i) selecting representative contracts with clustering method, (ii) pricing these contracts with Monte Carlo (MC) simulation, and (iii) predicting the value of the whole portfolio based on the values of representative contracts with kriging method. To further boost the efficiency and the effectiveness of selecting and pricing the representative contracts, as well as valuating the whole portfolio, various methods at each of these three steps have been proposed. For instance, Gan & Lin (2015) extended the ordinary kriging method to the universal kriging method; Hejazi & Jackson (2016) used a neural network as the predictive model to valuate the whole portfolio; Gan & Valdez (2018) implemented the generalised beta of the second kind method instead of the kriging method to capture the non-Gaussian behaviour of the market price of variable annuities. See also, Gan (2018), Gan & Valdez (2020), Gweon *et al.* (2020), Liu & Tan (2020), Lin & Yang (2020), Feng *et al.* (2020), and Quan *et al.* (2021) for recent developments in this three-step technique. Similar idea has also been applied to the calculation of Greeks and risk measures of a portfolio of variable annuities; see Gan & Lin (2017), Gan & Valdez (2017), and Xu *et al.* (2018). All of the above literature applying the machine learning methods involve the supervised learning, which requires a pre-labelled dataset (in this case, it is the set of fair prices of the representative contracts) to train a predictive model.

Other than valuating and hedging variable annuities, supervised learning methods have also been applied to different actuarial contexts. Wüthrich (2018) used a neural network for the chain-ladder factors in the chain-ladder claim reserving model to include heterogeneous individual claim features. Gao & Wüthrich (2019) applied a convolutional neural network to classify drivers using their telematics data. Cheridito *et al.* (2020) estimated the risk measures of a portfolio of assets and liabilities with a feedforward neural network. Richman & Wüthrich (2021) and Perla *et al.* (2021) studied the mortality rate forecasting problem, where Richman & Wüthrich (2021) extended the traditional Lee-Carter model to multiple populations using a neural network, while Perla *et al.* (2021) applied deep learning techniques directly on a time-series data of mortality rate. Hu *et al.* (2022) modified the loss function in tree-based models to improve the predictive performance when applying to imbalanced datasets which are common in the insurance practice.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Meanwhile, a flourishing sub-field in machine learning, called the reinforcement learning (RL), has been skyrocketing and has proved its powerfulness in various tasks; see Silver *et al.* (2017), and the references therein. Contrary to the supervised learning, the RL does not require a pre-labelled dataset for training. Instead, in the RL, an *agent interacts* with an *environment*, by sequentially *observing states*, *taking*, as well as *revising*, *actions*, and *collecting rewards*. Without possessing any prior knowledge of the environment, the agent needs to *explore* the environment while *exploit* the collected reward signals, for learning. For a representative monograph of RL, see Sutton & Barto (2018); for its broad applications in economics, game theory, operations research, and finance, see the recent survey paper by Charpentier *et al.* (2021).

The mechanism of RL resembles how a hedging agent hedges any contingent claim dynamically. Indeed, the hedging agent could not know any specifics of the market environment, but could only observe states from the environment[1], take a hedging strategy, and learn from reward signals to progressively improve the hedging strategy. However, in the context of hedging, if an insurer builds a hedging agent based on a certain RL method, called RL agent hereafter, and allows this infant RL agent to interact and learn from the market environment right away, the insurer could bear enormous financial loss while the infant RL agent is still exploring the environment before it could effectively exploit the reward signals. Moreover, provided that the insurer could not know any specifics of the market environment as well, they could not supply any information derived from theoretical models to the infant RL agent, and thus, the agent could only obtain the reward signals via the realised terminal profit and loss, based on the realised net liability and hedging portfolio value; these signals should not be effective for an infant RL agent to learn from the market environment.

To resolve these two issues above, we propose a *two-phase (deep) RL approach*, which is composed of a *training phase* and an *online learning phase*. In the training phase, based on their best knowledge of the market, the insurer constructs a training environment. An infant RL agent is then designated to interact and learn from this training environment for a period of time. Comparing to putting the infant RL agent in the market environment right away, the infant RL agent could be supplied by more information derived from the constructed training environment, such as the net liabilities before any terminal times. In this paper, we propose that the RL agent collects *anchor-hedging reward signals* during the training phase. After the RL agent is experienced with the training environment, in the online learning phase, the insurer finally designates the trained RL agent in the market environment. Again, since no theoretical model for the market environment is available to the insurer, the trained RL agent could only collect *single terminal reward signals* in this phase. In this paper, an illustrative example is provided to demonstrate the hedging performance using this approach.

All RL methods can be classified into either MC or temporal-difference (TD) learning. As a TD method shall be employed in this paper, in both the training and online learning phases, the following RL literature review focuses on the latter method. Sutton (1984) and (1988) first introduced the TD method for prediction of value function. Based upon their works, Watkins (1989) and Watkins & Dayan (1992) proposed the well-known Q-learning for finite state and action spaces. Since then, the Q-learning has been improved substantially, in Hasselt (2010) for the Double Q-learning, and in Mnih *et al.* (2013), as well as Mnih *et al.* (2015), for the deep Q-learning which allows infinite state space. Any Q-learning approaches, or in general tabular solution methods and value function approximation methods, are only applicable to finite action space. However, in the context of hedging, the action space is infinite. Instead of discretising the action space, *proximal policy optimisation* (PPO) by Schulman *et al.* (2017), which is a *policy gradient method*, shall be applied in this paper; our section 3.4 shall provide its self-contained review.

[1]Note that a "state" in this paper, in line with the terminologies of Markov decision processes, refers to observable metrics from the environment to be a proxy of a true state with unobservable but desirable features. See sections 2.3.2 and 7.1 for more details.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

To the best of our knowledge, this paper is the first work to implement the RL algorithms with online learning to hedge contingent claims, particularly variable annuities. Contrary to Xu (2020) and Carbonneau (2021), in which both adapted the state-of-the-art DH approach in Bühler *et al.* (2019), this paper is in line with the recent works by Kolm & Ritter (2019) and Cao *et al.* (2021), while extends with actuarial components. We shall outline the differences between the RL and DH approaches throughout sections 3 and 4, as well as Appendices A and B. Kolm & Ritter (2019) discretised the action space and implemented RL algorithms for finitely many possible actions; however, as mentioned above, this paper does not discretise the action space but adapts the recently advanced policy gradient method, namely, the PPO. Comparing with Cao *et al.* (2021), in addition to the actuarial elements, this paper puts forward online learning to self-revise the hedging strategy.

In the illustrative example, we assume that the market environment is the BS financial and constant force of mortality (CFM) actuarial markets, and the focus is on contracts with both GMMB and GMDB riders. Furthermore, we assume that the model of the market environment being presumed by the insurer, which shall be supplied as the training environment, is also the BS and the CFM, but with a different set of parameters. That is, while the insurer constructs correct dynamic models of the market environment for the training environment, the parameters in the model of the market environment are not the same as those in the market environment. Section 2.4 shall set the stage of this illustrative example and shall show that, if the insurer forwardly implements, in the market environment, the incorrect Delta hedging strategy based on their presumed model of the market environment, then its hedging performance for the variable annuities is worse than that by the correct Delta hedging strategy based on the market environment. In sections 4 and 6, this illustrative example shall be revisited using the two-phase RL approach. As we shall see in section 6, the hedging performance of the RL agent is even worse than that of the incorrect Delta, at the very beginning of hedging in real time. However, delicate analysis shows that, with a fair amount of future trajectories (which are different from simulated scenarios, with more details in section 6), the hedging performance of the RL agent becomes comparable with that of the correct Delta within a reasonable amount of time. Therefore, the illustrative example addresses model miscalibration issue in hedging variable annuity contracts with GMMB and GMDB riders in BS financial and CFM actuarial market environments, which is common in practice.

This paper is organised as follows. Section 2 formulates the continuous hedging problem for variable annuities, reformulates it to the discrete and Markov setting, and motivates as well as outlines the two-phase RL approach. Section 3 discusses the RL approach in hedging variable annuities and provides a self-contained review of RL, particularly the PPO, which is a TD policy gradient method, while section 5 presents the implementation details of the online learning phase. Sections 4 and 6 revisit the illustrative example in the training and online learning phases, respectively. Section 7 collates the assumptions of utilising the two-phase RL approach for hedging contingent claims, as well as their implications in practice. This paper finally concludes and comments on future directions in section 8.

2. Problem Formulation and Motivation

2.1. Classical hedging problem and model-based approach

We first review the classical hedging problem for variable annuities and its model-based solution to introduce some notations and to motivate the RL approach.

2.1.1. Actuarial and financial market models

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a rich enough complete probability space. Consider the current time $t = 0$ and fix $T > 0$ as a deterministic time in the future. Throughout this paper, all time units are in year.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

There are one risk-free asset and one risky asset in the financial market. Let B_t and S_t, for $t \in [0, T]$, be the time-t values of the risk-free asset and the risky asset, respectively. Let $\mathbb{G}^{(1)} = \left\{\mathcal{G}_t^{(1)}\right\}_{t\in[0,T]}$ be the filtration which contains all financial market information; in particular, both processes $B = \{B_t\}_{t\in[0,T]}$ and $S = \{S_t\}_{t\in[0,T]}$ are $\mathbb{G}^{(1)}$-adapted.

There are N policyholders in the actuarial market. For each policyholder $i = 1, 2, \ldots, N$, denote $T_{x_i}^{(i)}$ as their random future lifetime, who is of age x_i at the current time 0. Define, for each $i = 1, 2, \ldots, N$, and for any $t \geq 0$, $J_t^{(i)} = \mathbb{1}_{\left\{T_{x_i}^{(i)} > t\right\}}$, be the corresponding time-t jump value generated by the random future lifetime of the i-th policyholder; that is, if the i-th policyholder survives at some time $t \in [0, T]$, $J_t^{(i)} = 1$; otherwise, $J_t^{(i)} = 0$. Let $\mathbb{G}^{(2)} = \left\{\mathcal{G}_t^{(2)}\right\}_{t\in[0,T]}$ be the filtration which contains all actuarial market information; in particular, all single-jump processes $J^{(i)} = \left\{J_t^{(i)}\right\}_{t\in[0,T]}$, for $i = 1, 2, \ldots, N$, are $\mathbb{G}^{(2)}$-adapted.

Let $\mathbb{F} = \{\mathcal{F}_t\}_{t\in[0,T]}$ be the filtration which contains all actuarial and financial market information; that is, $\mathbb{F} = \mathbb{G}^{(1)} \vee \mathbb{G}^{(2)}$. Therefore, the filtered probability space is given by $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$.

2.1.2. Variable annuities with guaranteed minimum maturity benefit and guaranteed minimum death benefit riders

At the current time 0, an insurer writes a variable annuity contract to each of these N policyholders. Each contract is embedded with both GMMB and GMDB riders. Assume that all these N contracts expire at the same fixed time T. In the following, fix a generic policyholder $i = 1, 2, \ldots, N$.

At the current time 0, the policyholder deposits $F_0^{(i)}$ into their segregated account to purchase $\rho^{(i)} > 0$ shares of the risky asset; that is, $F_0^{(i)} = \rho^{(i)} S_0$. Assume that the policyholder does not revise the number of shares $\rho^{(i)}$ throughout the effective time of the contract.

For any $t \in \left[0, T_{x_i}^{(i)} \wedge T\right]$, the time-$t$ segregated account value of the policyholder is given by $F_t^{(i)} = \rho^{(i)} S_t e^{-m^{(i)}t}$, where $m^{(i)} \in (0, 1)$ is the continuously compounded annualised rate at which the asset-value-based fees are deducted from the segregated account by the insurer. For any $t \in \left(T_{x_i}^{(i)} \wedge T, T\right]$, the time-$t$ segregated account value $F_t^{(i)}$ must be 0; indeed, if the policyholder dies before the maturity, i.e. $T_{x_i}^{(i)} < T$, then, due to the GMDB rider of a minimum guarantee $G_D^{(i)} > 0$, the beneficiary inherits $\max\left\{F_{T_{x_i}^{(i)}}^{(i)}, G_D^{(i)}\right\}$, which can be decomposed into $F_{T_{x_i}^{(i)}}^{(i)} + \left(G_D^{(i)} - F_{T_{x_i}^{(i)}}^{(i)}\right)_+$, at the policyholder's death time $T_{x_i}^{(i)}$ right away. Due to the GMMB rider of a minimum guarantee $G_M^{(i)} > 0$, if the policyholder survives beyond the maturity, i.e. $T_{x_i}^{(i)} > T$, the policyholder acquires $\max\left\{F_T^{(i)}, G_M^{(i)}\right\}$ at the maturity, which can be decomposed into $F_T^{(i)} + \left(G_M^{(i)} - F_T^{(i)}\right)_+$.

2.1.3. Net liability of insurer

The liability of the insurer thus has two parts. The liability from the GMMB rider at the maturity for the i-th policyholder, where $i = 1, 2, \ldots, N$, is given by $\left(G_M^{(i)} - F_T^{(i)}\right)_+$ if the i-th policyholder survives beyond the maturity, and is 0 otherwise. The liability from the GMDB rider at the death time $T_{x_i}^{(i)}$ for the i-th policyholder, where $i = 1, 2, \ldots, N$, is given by $\left(G_D^{(i)} - F_{T_{x_i}^{(i)}}^{(i)}\right)_+$ if the i-th policyholder dies before the maturity, and is 0 otherwise. Therefore, at any time $t \in [0, T]$, the

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

future gross liability of the insurer accumulated to the maturity for these N contracts is given by

$$\sum_{i=1}^{N}\left(\left(G_M^{(i)}-F_T^{(i)}\right)_+ J_T^{(i)}+\frac{B_T}{B_{T_{x_i}^{(i)}}}\left(G_D^{(i)}-F_{T_{x_i}^{(i)}}^{(i)}\right)_+ \mathbb{1}_{\{T_{x_i}^{(i)}<T\}}J_t^{(i)}\right).$$

Denote V_t^{GL}, for $t\in[0,T]$, as the time-t value of the discounted (via the risk-free asset B) future gross liability of the insurer; if the liability is 0, the value will be 0.

From the asset-value-based fees collected by the insurer, a portion, known as the rider charge, is used to fund the liability due to the GMMB and GMDB riders; the remaining portion is used to cover overhead, commissions, and any other expenses. From the i-th policyholder, where $i=1,2,\ldots,N$, the insurer collects $m_e^{(i)}F_t^{(i)}J_t^{(i)}$ as the rider charge at any time $t\in[0,T]$, where $m_e^{(i)}\in\left(0,m^{(i)}\right]$. Therefore, the cumulative future rider charge to be collected, from any time $t\in[0,T]$ onward, till the maturity, by the insurer from these N policyholders, is given by $\sum_{i=1}^{N}\int_t^T m_e^{(i)}F_s^{(i)}J_s^{(i)}(B_T/B_s)\,ds$. Denote V_t^{RC}, for $t\in[0,T]$, as its time-t discounted (via the risk-free asset B) value; if the cumulative rider charge is 0, the value will be 0.

Hence, due to these N variable annuity contracts with both GMMB and GMDB riders, for any $t\in[0,T]$, the time-t net liability of the insurer for these N contracts is given by $L_t=V_t^{\mathrm{GL}}-V_t^{\mathrm{RC}}$, which is $\mathcal{F}_t$-measurable.

One of the many ways to set the rate $m^{(i)}\in(0,1)$ for the asset-value-based fees, and the rate $m_e^{(i)}\in\left(0,m^{(i)}\right]$ for the rider charge, for $i=1,2,\ldots,N$, is based on the time-0 net liability of the insurer for the i-th policyholder. More precisely, $m^{(i)}$ and $m_e^{(i)}$ are determined via $L_0^{(i)}=V_0^{\mathrm{GL},(i)}-V_0^{\mathrm{RC},(i)}=0$, where $V_0^{\mathrm{GL},(i)}$ and $V_0^{\mathrm{RC},(i)}$ are the time-0 values of, respectively, the discounted future gross liability and the discounted cumulative future rider charge, of the insurer for the i-th policyholder.

2.1.4. Continuous hedging and hedging objective

The insurer aims to hedge this dual-risk bearing net liability via investing in the financial market. To this end, let $\tilde{T}$ be the death time of the last policyholder; that is, $\tilde{T}=\max_{i=1,2,\ldots,N}T_{x_i}^{(i)}$, which is random.

While the net liability L_t is defined for any time $t\in[0,T]$, as the difference between the values of discounted future gross liability and discounted cumulative future rider charge, $L_t=0$ for any $t\in\left(\tilde{T}\wedge T,T\right]$. Indeed, if $\tilde{T}<T$, then, for any $t\in\left(\tilde{T}\wedge T,T\right]$, one has $T_{x_i}^{(i)}<t\le T$ for all $i=1,2,\ldots,N$, and hence, the future gross liability accumulated to the maturity, and the cumulative rider charge from time $\tilde{T}$ onward are both 0 so are their values. Therefore, the insurer only hedges the net liability L_t, for any $t\in\left[0,\tilde{T}\wedge T\right]$.

Let H_t be the hedging strategy, i.e. the number of shares of the risky asset being held by the insurer, at time $t\in[0,T)$. Hence, $H_t=0$, for any $t\in\left[\tilde{T}\wedge T,T\right)$. Let $\mathcal{H}$ be the admissible set of hedging strategies, which is defined by

$$\mathcal{H}=\Big\{H=\{H_t\}_{t\in[0,T)}\ :\text{(i) } H \text{ is } \mathbb{F}\text{-adapted, (ii) } H\in\mathbb{R},\ \mathbb{P}\times\mathcal{L}\text{-a.s., and}$$
$$\text{(iii) for any } t\in\left[\tilde{T}\wedge T,T\right),\ H_t=0\Big\},$$

where $\mathcal{L}$ is the Lebesgue measure on $\mathbb{R}$. The condition (ii) indicates that there is not any constraint on the hedging strategies.

Let P_t be the time-t value, for $t\in[0,T]$, of the insurer's hedging portfolio. Then, $P_0=0$, and together with the rider charges collected from the N policyholders, as well as the withdrawal

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

for paying the liabilities due to the beneficiaries' inheritance from those policyholders who have already been dead, for any $t \in (0, T]$,

$$P_t = \int_0^t (P_s - H_s S_s) \frac{dB_s}{B_s} + \int_0^t H_s dS_s + \sum_{i=1}^{N} \int_0^t m_e^{(i)} F_s^{(i)} J_s^{(i)} ds - \sum_{i=1}^{N} \left(G_D^{(i)} - F_{T_{x_i}^{(i)}}^{(i)} \right)_+ \mathbb{1}_{\{T_{x_i}^{(i)} \le t < T\}},$$

which obviously depends on $\{H_s\}_{s\in[0,t)}$.

As in Bertsimas *et al.* (2000), the insurer's hedging objective function at the current time 0 should be given by the root-mean-square error (RMSE) of the terminal profit and loss (P&L), which is, for any $H \in \mathcal{H}$,

$$\sqrt{\mathbb{E}^{\mathbb{P}}\left[\left(P_{\tilde{T}\wedge T} - L_{\tilde{T}\wedge T}\right)^2\right]}.$$

If the insurer has full knowledge of the objective probability measure $\mathbb{P}$, and hence the correct dynamics of the risk-free asset and the risky asset in the financial market, as well as the correct mortality model in the actuarial market, the optimal hedging strategy, being implemented forwardly, is given by minimising the RMSE of the terminal P&L:

$$H^* = \arg\min_{H\in\mathcal{H}} \sqrt{\mathbb{E}^{\mathbb{P}}\left[\left(P_{\tilde{T}\wedge T} - L_{\tilde{T}\wedge T}\right)^2\right]}.$$

2.2. Pitfall of model-based approach

However, having correct model is usually not the case in practice. Indeed, the insurer, who is the hedging agent above, usually has little information regarding the objective probability measure $\mathbb{P}$ and hence easily misspecifies the financial market dynamics and the mortality model, which will in turn yield a poor performance from the supposedly optimal hedging strategy when it is implemented forwardly in the future. Section 2.4 outlines such an illustrative example which shall be discussed throughout the remaining of this paper.

To rectify this, we propose a *two-phase (deep) RL approach* to solve an optimal hedging strategy. In this approach, an RL agent, which is not the insurer themselves but is built by the insurer to hedge on their behalf, does not have any knowledge of the objective probability measure $\mathbb{P}$, the financial market dynamics, and the mortality model; section 2.5 shall explain this approach in details. Before that, in the following section 2.3, the classical hedging problem shall first be reformulated with a Markov decision process (MDP) in a discrete time setting so that RL methods can be implemented. The illustrative example outlined in section 2.4 shall be revisited using the proposed two-phase RL approach in sections 4 and 6.

In the remaining of this paper, unless otherwise specified, all expectation operators shall be taken with respect to the objective probability measure $\mathbb{P}$ and denoted simply as $\mathbb{E}[\cdot]$.

2.3. Discrete and Markov hedging

2.3.1. Discrete hedging and hedging objective

Let $t_0, t_1, \ldots, t_{n-1} \in [0, T)$, for some $n \in \mathbb{N}$, be the time when the hedging agent decides the hedging strategy, such that $0 = t_0 < t_1 < \cdots < t_{n-1} < T$. Denote also $t_n = T$.

Let $t_{\tilde{n}}$ be the first time (right) after the last policyholder dies or all contracts expire, for some $\tilde{n} = 1, 2, \ldots, n$, which is random; that is, $t_{\tilde{n}} = \min\left\{t_k,\ k = 1, 2, \ldots, n : t_k \ge \tilde{T}\right\}$, and when $\tilde{T} > T$, by convention, $\min \emptyset = t_n$. Therefore, $H_t = 0$, for any $t = t_{\tilde{n}}, t_{\tilde{n}+1}, \ldots, t_{n-1}$. With a slight abuse of

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

notation, the admissible set of hedging strategies in discrete time is

$$\mathcal{H} = \Big\{H = \{H_t\}_{t=t_0,t_1,\ldots,t_{n-1}} : \text{(i) for any } t = t_0, t_1, \ldots, t_{n-1},\ H_t \text{ is } \mathcal{F}_t\text{-measurable,}$$
$$\text{(ii) for any } t = t_0, t_1, \ldots, t_{n-1},\ H_t \in \mathbb{R},\ \mathbb{P}\text{-a.s., and}$$
$$\text{(iii) for any } t = t_{\tilde{n}}, t_{\tilde{n}+1}, \ldots, t_{n-1},\ H_t = 0\Big\};$$

again, the condition (ii) emphasises that no constraint is imposed to the hedging strategies.

While the hedging agent decides the hedging strategy at the discrete time points, the actuarial and financial market models are continuous. Hence, the net liability $L_t = V_t^{\mathrm{GL}} - V_t^{\mathrm{RC}}$ is still defined for any time $t \in [0, T]$ as before. Moreover, if $t \in \big[t_k, t_{k+1}\big)$, for some $k = 0, 1, \ldots, n-1$, $H_t = H_{t_k}$; thus, $P_0 = 0$, and, if $t \in \big(t_k, t_{k+1}\big]$, for some $k = 0, 1, \ldots, n-1$,

$$P_t = \big(P_{t_k} - H_{t_k} S_{t_k}\big)\frac{B_t}{B_{t_k}} + H_{t_k} S_t + \sum_{i=1}^{N} \int_{t_k}^{t} m_e^{(i)} F_s^{(i)} J_s^{(i)} \frac{B_t}{B_s} ds$$
$$- \sum_{i=1}^{N} \frac{B_t}{B_{T_{x_i}^{(i)}}} \Big(G_D^{(i)} - F_{T_{x_i}^{(i)}}^{(i)}\Big)_+ \mathbb{1}_{\{t_k < T_{x_i}^{(i)} \le t < T\}}. \tag{1}$$

For any $H \in \mathcal{H}$, the hedging objective of the insurer at the current time 0 is $\sqrt{\mathbb{E}\Big[\big(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\big)^2\Big]}$. Hence, the optimal discrete hedging strategy, being implemented forwardly, is given by

$$H^* = \arg\min_{H \in \mathcal{H}} \sqrt{\mathbb{E}\Big[\big(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\big)^2\Big]} = \arg\min_{H \in \mathcal{H}} \mathbb{E}\Big[\big(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\big)^2\Big]. \tag{2}$$

2.3.2. Markov decision process

An MDP can be characterised by its state space, action space, Markov transition probability, and reward signal. In turn, these derive the value function and the optimal value function, which are equivalently known as, respectively, the objective function and the value function, in optimisation as in the previous sections. In the remaining of this paper, we shall adapt the MDP language.

- (*State*) Let $\mathcal{X}$ be the state space in $\mathbb{R}^p$, where $p \in \mathbb{N}$. Each state in the state space represents a possible observation with p features in the actuarial and financial markets. Denote $X_{t_k} \in \mathcal{X}$ as the observed state at any time t_k, where $k = 0, 1, \ldots, n$; the state should minimally include an information related to the number of surviving policyholders $\sum_{i=1}^{N} J_{t_k}^{(i)}$, and the term to maturity $T - t_k$, in order to terminate the hedging at time $t_{\tilde{n}}$, which is the first time when $\sum_{i=1}^{N} J_{t_{\tilde{n}}}^{(i)} = 0$, or which is when $T - t_{\tilde{n}} = 0$. The states (space) shall be specified in sections 4 and 5.
- (*Action*) Let $\mathcal{A}$ be the action space in $\mathbb{R}$. Each action in the action space is a possible hedging strategy. Denote $H_{t_k}\big(X_{t_k}\big) \in \mathcal{A}$ as the action at any time t_k, where $k = 0, 1, \ldots, n-1$, which is assumed to be Markovian with respect to the observed state X_{t_k}; that is, given the current state X_{t_k}, the current action $H_{t_k}\big(X_{t_k}\big)$ is independent of the past states $X_{t_0}, X_{t_1}, \ldots, X_{t_{k-1}}$. In the sequel, for notational simplicity, we simply write H_{t_k} to represent $H_{t_k}\big(X_{t_k}\big)$, for $k = 0, 1, \ldots, n-1$. If the feature of the number of surviving policyholders $\sum_{i=1}^{N} J_{t_k}^{(i)} = 0$, for $k = 0, 1, \ldots, n-1$, in the state X_{t_k}, then $H_{t_k} = 0$; in particular, for any t_k, where $k = \tilde{n}, \tilde{n}+1, \ldots, n-1$, the hedging strategy $H_{t_k} = 0$.
- (Markov property) At any time t_k, where $k = 0, 1, \ldots, n-1$, given the current state X_{t_k} and the current hedging strategy H_{t_k}, the transition probability distribution of the next state

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

$X_{t_{k+1}}$ in the market is independent of the past states $X_{t_0}, X_{t_1}, \ldots, X_{t_{k-1}}$ and the past hedging strategies $H_{t_0}, H_{t_1}, \ldots, H_{t_{k-1}}$; that is, for any Borel set $\overline{B} \in \mathcal{B}(\mathcal{X})$,

$$\mathbb{P}\left(X_{t_{k+1}} \in \overline{B} | H_{t_k}, X_{t_k}, H_{t_{k-1}}, X_{t_{k-1}}, \ldots, H_{t_1}, X_{t_1}, H_{t_0}, X_{t_0}\right) = \mathbb{P}\left(X_{t_{k+1}} \in \overline{B} | H_{t_k}, X_{t_k}\right). \quad (3)$$

- (*Reward*) At any time t_k, where $k = 0, 1, \ldots, n-1$, given the current state X_{t_k} in the market and the current hedging strategy H_{t_k}, a reward signal $R_{t_{k+1}}\left(X_{t_k}, H_{t_k}, X_{t_{k+1}}\right)$ is received, by the hedging agent, as a result of transition to the next state $X_{t_{k+1}}$. The reward signal shall be specified after introducing the (optimal) value function below. In the sequel, occasionally, for notational simplicity, we simply write $R_{t_{k+1}}$ to represent $R_{t_{k+1}}\left(X_{t_k}, H_{t_k}, X_{t_{k+1}}\right)$, for $k = 0, 1, \ldots, n-1$.
- (State, action, and reward sequence) The states, actions (which are hedging strategies herein), and reward signals form an *episode*, which is sequentially given by:

$$\left\{X_{t_0}, H_{t_0}, X_{t_1}, R_{t_1}, H_{t_1}, X_{t_2}, R_{t_2}, H_{t_2}, \ldots, X_{t_{\tilde{n}-1}}, R_{t_{\tilde{n}-1}}, H_{t_{\tilde{n}-1}}, X_{t_{\tilde{n}}}, R_{t_{\tilde{n}}}\right\}.$$

- (Optimal value function) Based on the reward signals, the value function, at any time t_k, where $k = 0, 1, \ldots, n-1$, with the state $x \in \mathcal{X}$, is defined by, for any hedging strategies $H_{t_k}, H_{t_{k+1}}, \ldots, H_{t_{n-1}}$,

$$V\left(t_k, x; H_{t_k}, H_{t_{k+1}}, \ldots, H_{t_{n-1}}\right) = \mathbb{E}\left[\sum_{l=k}^{n-1} \gamma^{t_{l+1}-t_k} R_{t_{l+1}} \middle| X_{t_k} = x\right], \quad (4)$$

where $\gamma \in [0, 1]$ is the discount rate; the value function, at the time $t_n = T$ with the state $x \in \mathcal{X}$, is defined by $V(t_n, x) = 0$. Hence, the optimal discrete hedging strategy, being implemented forwardly, is given by

$$H^* = \arg\max_{H \in \mathcal{H}} \mathbb{E}\left[\sum_{k=0}^{n-1} \gamma^{t_{k+1}} R_{t_{k+1}} \middle| X_0 = x\right]. \quad (5)$$

In turn, the optimal value function, at any time t_k, where $k = 0, 1, \ldots, n-1$, with the state $x \in \mathcal{X}$, is

$$V^*(t_k, x) = V\left(t_k, x; H^*_{t_k}, H^*_{t_{k+1}}, \ldots, H^*_{t_{n-1}}\right), \text{ and } V^*(t_n, x) = 0. \quad (6)$$

- (Reward engineering) To ensure the hedging problem being reformulated with the MDP, the value functions, given by that in (5), and the negative of that in (2), should coincide; that is,

$$\mathbb{E}\left[\sum_{k=0}^{n-1} \gamma^{t_{k+1}} R_{t_{k+1}} \middle| X_0 = x\right] = -\mathbb{E}\left[\left(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\right)^2\right]. \quad (7)$$

Hence, two possible constructions for the reward signals are proposed as follows; each choice of the reward signals shall be utilised in one of the two phases in the proposed RL approach.

  - (Single terminal reward) An obvious choice is to only have a reward signal from the negative squared terminal P&L; that is, for any time t_k,

$$R_{t_{k+1}} = \begin{cases} -\left(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\right)^2 & \text{if } k = \tilde{n} - 1, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

    Necessarily, the discount rate is given as $\gamma = 1$.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 1. Contract characteristics.

Parameter	Value
Expiration date T	1
Minimum guarantee at maturity G_M	100
Minimum guarantee at death G_D	100

– (Sequential anchor-hedging reward) A less obvious choice is via telescoping the RHS of Equation (7), that

$$-\mathbb{E}\left[\left(P_{t_{\tilde{n}}}-L_{t_{\tilde{n}}}\right)^2\right]=-\mathbb{E}\left[\sum_{k=0}^{\tilde{n}-1}\left(\left(P_{t_{k+1}}-L_{t_{k+1}}\right)^2-\left(P_{t_k}-L_{t_k}\right)^2\right)+\left(P_0-L_0\right)^2\right].$$

Therefore, when $L_0 = P_0$, another possible construction for the reward signal is, for any time t_k,

$$R_{t_{k+1}}=\begin{cases}\left(P_{t_k}-L_{t_k}\right)^2-\left(P_{t_{k+1}}-L_{t_{k+1}}\right)^2 & \text{if } k=0,1,\ldots,\tilde{n}-1,\\ 0 & \text{otherwise.}\end{cases} \tag{9}$$

Again, the discount rate is necessarily given as $\gamma = 1$. The constructed reward in (9) outlines an *anchor-hedging* scheme. First, note that, at the current time 0, when $L_0 = P_0$, there is no local hedging error. Then, at each future hedging time before the last policyholder dies and before the maturity, the hedging performance is measured by the local squared P&L, i.e. $\left(P_{t_k}-L_{t_k}\right)^2$, which serves as an anchor. At the next hedging time, if the local squared P&L is smaller than the anchor, it will be rewarded, i.e. $R_{t_{k+1}} > 0$; however, if the local squared P&L becomes larger, it will be penalised, i.e. $R_{t_{k+1}} < 0$.

2.4. Illustrative example

The illustrative example below demonstrates the poor hedging performance by the Delta hedging strategy when the insurer miscalibrates the parameters in the market environment. We consider that the insurer hedges a variable annuity contract, with both GMMB and GMDB riders, of a single policyholder, i.e. $N = 1$, with the contract characteristics given in Table 1.

The market environment follows the Black-Scholes (BS) in the financial part and the constant force of mortality (CFM) in the actuarial front. The risk-free asset earns a constant risk-free interest rate $r > 0$ that, for any $t \in [0, T]$, $dB_t = rB_t dt$, while the value of the risky asset evolves as a geometric Brownian motion that, for any $t \in [0, T]$, $dS_t = \mu S_t dt + \sigma S_t dW_t$, where μ is a constant drift, $\sigma > 0$ is a constant volatility, and $W = \{W_t\}_{t\in[0,T]}$ is the standard Brownian motion. The random future lifetime of the policyholder T_x has a CFM $\nu > 0$; that is, for any $0 \le t \le s \le T$, the conditional survival probability $\mathbb{P}(T_x, > s|T_x > t) = e^{-\nu(s-t)}$. Moreover, the Brownian motion W in the financial market and the future lifetime T_x in the actuarial market are independent. Table 2 summarises the parameters in the market environment. Note that the risk-free interest rate, the risky asset initial price, the initial age of the policyholder, and the investment strategy of the policyholder are observable by the insurer.

Based on their best knowledge of the market, the insurer builds a model of the market environment. Suppose that the model happens to be the BS and the CFM as the market environment, but the insurer *miscalibrates* the parameters. Table 3 lists these parameters in the model of the market environment. In particular, the risky asset drift and volatility, as well as the force of mortality constant, are different from those in the market environment. For the observable parameters, they are the same as those in the market environment.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 2. Parameters setting of market environment.

(a) Black-Scholes financial market	
Parameter	Value
Risk-free interest rate r	0.02
Risky asset initial price S_0	100
Risky asset drift μ	-0.2
Risky asset volatility σ	0.4
(b) Constant force of mortality actuarial market	
Parameter	Value
Initial number of policyholders N	1
Initial age of policyholders x	20
Constant force of mortality ν	0.03
Investment strategy of policyholders ρ	1.19

Table 3. Parameters setting of model of market environment, with bolded parameters being different from those in market environment.

(a) Black-Scholes financial market	
Parameter	Value
Risk-free interest rate r	0.02
Risky asset initial price S_0	100
Risky asset drift μ	0.08
Risky asset volatility σ	0.2
(b) Constant force of mortality actuarial market	
Parameter	Value
Initial number of policyholders N	1
Initial age of policyholders x	20
Constant force of mortality ν	0.02
Investment strategy of policyholders ρ	1.19

At any time $t \in [0, T]$, the value of the hedging portfolio of the insurer is given by (17), with $N = 1$, in which the values of the risky asset and the single-jump process follow the market environment with the parameters in Table 2. At any time $t \in [0, T]$, the value of the net liability of the insurer is given by (16), with $N = 1$, in both the market environment and its model; for its detailed derivations, we defer it to section 4.1, as the model of the market environment, with multiple homogeneous policyholders for effective training, shall be supplied as the training environment. Since the parameters in the model of the market environment (see Table 3) are different from those in the market environment (see Table 2), the net liability evaluated by the insurer using the model is different from that of the market environment. There are two implications. Firstly, the Delta hedging strategy of the insurer using the parameters in Table 3 is incorrect, while the correct Delta hedging strategy should use the parameters in Table 2. Secondly, the asset-value-based fee m and the rider charge m_e given in Table 4, which are determined by the insurer based on the time-0 value of their net liability by Table 3 via the method in section 2.1.3, are mispriced. They would not lead to zero time-0 value of their net liability in the market environment which is based on Table 2.

To evaluate the hedging performance of the incorrect Delta strategy by the insurer in the market environment for the variable annuity of contract characteristics in Table 1, 5,000 market

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 4. Fee structures derived from model of market environment.

Parameter	Value
Rate for asset-value-based fee m	0.02
Rate for rider charge m_e	0.019

Table 5. Summary statistics of empirical distributions of realised terminal P&Ls by different Delta strategies.

Terminal P&L of hedging strategy	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}	$\widehat{\text{RMSE}}$
Correct Delta	−0.24	−0.14	2.96	−4.00	−5.59	−5.99	−7.22	2.97
Incorrect Delta	−1.25	−0.22	3.41	−6.27	−8.80	−9.24	−11.05	3.63

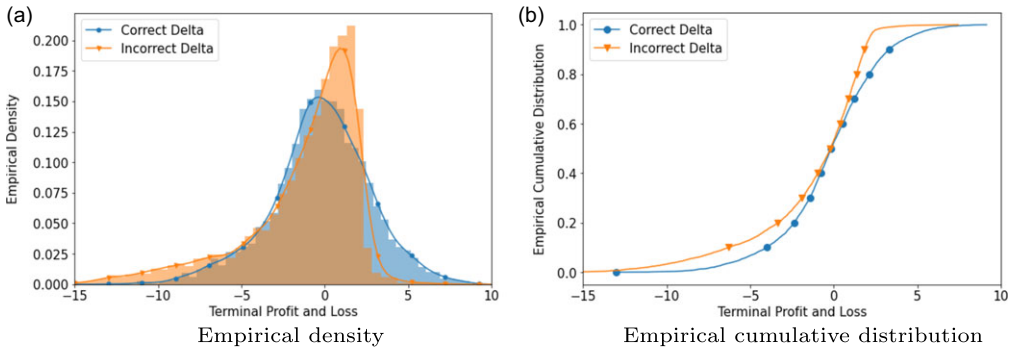


Figure 1. Empirical density and cumulative distribution functions of realised terminal P&Ls by different Delta strategies.

scenarios using the parameters in Table 2 are simulated to realise terminal P&Ls. For comparison, the terminal P&Ls by the correct Delta hedging strategy are also obtained. Figure 1 shows the empirical density and cumulative distribution functions of the 5,000 realised terminal P&Ls by each Delta hedging strategy, while Table 5 outlines the summary statistics of the empirical distributions, in which $\widehat{\text{RMSE}}$ is the estimated RMSE of the terminal P&L similar to (2).

In Figure 1(a), the empirical density function of realised terminal P&Ls by the incorrect Delta hedging strategy is depicted to be more heavy-tailed on the left than that by the correct Delta strategy. In fact, the terminal P&L by the incorrect Delta hedging strategy is stochastically dominated by that by the correct Delta strategy in the first-order; see Figure 1(b). Table 5 shows that the terminal P&L by the incorrect Delta hedging strategy has a mean and a median farther from zero, a higher standard deviation, larger left-tail risks in terms of Value-at-Risk and Tail Value-at-Risk, and a larger RMSE than that by the correct Delta strategy.

These observations conclude that, even in a market environment as simple as the BS and the CFM, the incorrect Delta hedging strategy based on the miscalibrated parameters by the insurer does not perform well when it is being implemented forwardly. In general, the hedging performance of model-based approaches depends crucially on the calibration of parameters for the model of the market environment.

2.5. Two-phase reinforcement learning approach

In an RL approach, at the current time 0, the insurer builds an RL agent to hedge on their behalf in the future. The *agent interacts* with a market *environment*, by sequentially *observing states*, *taking*, as well as *revising*, *actions*, which are the hedging strategies, and *collecting rewards*.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Without possessing any prior knowledge of the market environment, the agent needs to, *explore* the environment while *exploit* the collected reward signals, for effective learning.

An intuitive proposition would be allowing an infant RL agent to learn directly from such market environment, like the one in section 2.4, moving forward. However, recall that the insurer actually does not know any exact market dynamics in the environment and thus is not able to provide any theoretical model for the net liability to the RL agent. In turn, the RL agent could not receive any sequential anchor-hedging reward signal in (9) from the environment, but instead receives the single terminal reward signal in (8). Since the rewards, except the terminal one, are all zero, the infant RL agent would learn ineffectively from such sparse rewards, i.e. the RL agent shall take a tremendous amount of time to finally learn a nearly optimal hedging strategy in the environment. Most importantly, while the RL agent is exploring and learning from the environment, which is not a simulated one, the insurer could suffer from huge financial burden due to any sub-optimal hedging performances.

In view of this, we propose that the insurer should first designate the infant RL agent to interact and learn from a training environment, which is constructed by the insurer based on their best knowledge of the market, for example, the model of the market environment in section 2.4. Since the training environment is known to the insurer (but is unknown to the RL agent), the RL agent can be supplied by a net liability theoretical model, and consequently learn from the sequential anchor-hedging reward signal in (9) of the training environment. Therefore, the infant RL agent would be guided by the net liability to learn effectively from the local hedging errors. After interacting and learning from the training environment for a period of time, in order to gauge the effectiveness, the RL agent shall be tested for its hedging performance in simulated scenarios from the same training environment. This first phase is called the *training phase*.

Training Phase:

(i) The insurer constructs the MDP training environment.
(ii) The insurer builds the infant RL agent which uses the PPO algorithm.
(iii) The insurer assigns the RL agent in the MDP training environment to interact and learn for a period of time, during which the RL agent collects the anchor-hedging reward signal in (9).
(iv) The insurer deploys the trained RL agent to hedge in simulated scenarios from the same training environment and documents the baseline hedging performance.

If the hedging performance of the trained RL agent in the training environment is satisfactory, the insurer should then proceed to assign it to interact and learn from the market environment. Since the training and market environments are usually different, such as having different parameters as in section 2.4, the initial hedging performance of the trained RL agent in the market environment is expected to diverge from the fine baseline hedging performance in the training environment. However, different from an infant RL agent, the trained RL agent is experienced so that the sparse reward signal in (8) should be sufficient for the agent to revise the hedging strategy, from the nearly optimal one in the training environment to that in the market environment, within a reasonable amount of time. This second phase is called the *online learning phase*.

Online Learning Phase:

(v) The insurer assigns the RL agent in the market environment to interact and learn in real time, during which the RL agent collects the single terminal reward signal in (8).

These summarise the proposed two-phase RL approach. Figure 2 depicts the above sequence clearly. There are several assumptions underneath this two-phase RL approach in order to apply it effectively to a hedging problem of a contingent claim; as they involve specifics in later sections,

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

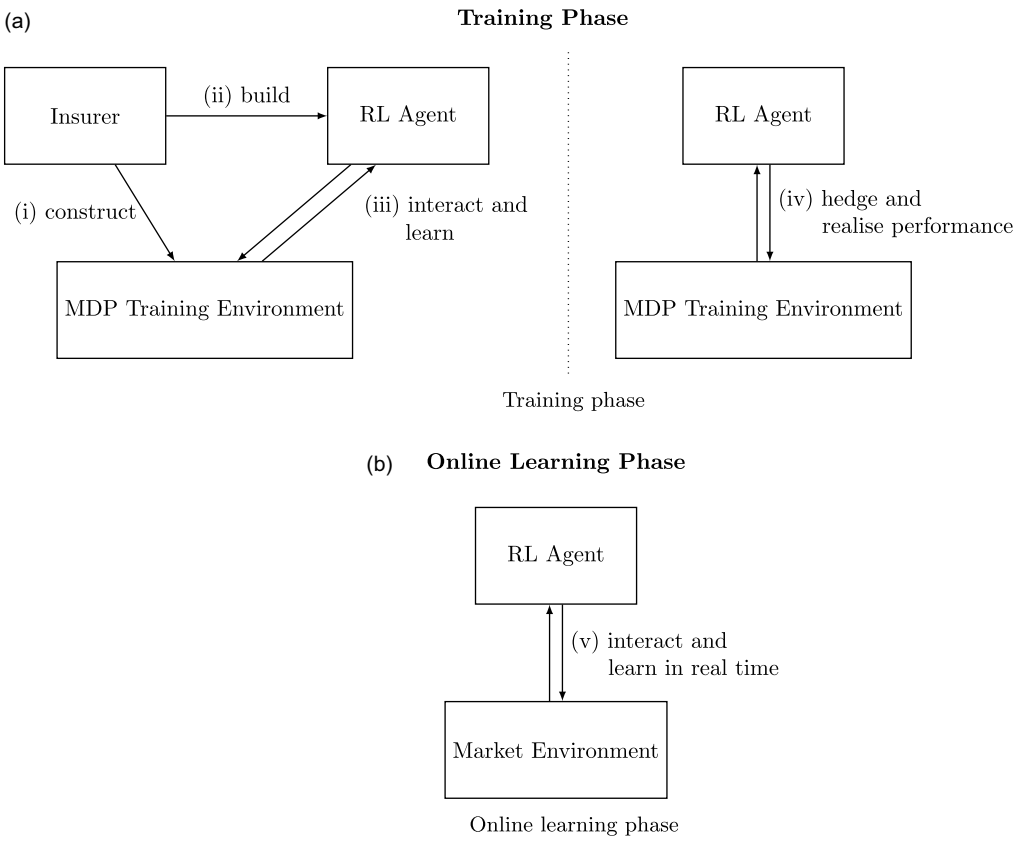


Figure 2. The relationship among insurer, RL agent, MDP training environment, and market environment of the two-phase RL approach.

we collate their discussions and elaborate their implications in practice in section 7. In the following section, we shall briefly review the training essentials of RL in order to introduce the PPO algorithm. For the details of online learning phase, we defer them until section 5.

3. Review of Reinforcement Learning

3.1. Stochastic action for exploration

One of the fundamental ideas in RL is that, at any time t_k, where $k = 0, 1, \ldots, n-1$, given the current state X_{t_k}, the RL agent does not take a deterministic action H_{t_k} but extends it to a stochastic action, in order to *explore* the MDP environment and in turn learn from the reward signals. The stochastic action is sampled through a so-called *policy*, which is defined below.

Let $\mathcal{P}(\mathcal{A})$ be a set of probability measures over the action space $\mathcal{A}$; each probability measure $\mu(\cdot) \in \mathcal{P}(\mathcal{A})$ maps a Borel set $\overline{A} \in \mathcal{B}(\mathcal{A})$ to $\mu\left(\overline{A}\right) \in [0, 1]$. The policy $\pi(\cdot)$ is a mapping from the state space $\mathcal{X}$ to the set of probability measures $\mathcal{P}(\mathcal{A})$; that is, for any state $x \in \mathcal{X}$, $\pi(x) = \mu(\cdot) \in \mathcal{P}(\mathcal{A})$. The value function and the optimal value function, at any time t_k, where $k = 0, 1, \ldots, \tilde{n}-1$, with the state $x \in \mathcal{X}$, are then generalised as, for any policy $\pi(\cdot)$,

$$V(t_k, x; \pi(\cdot)) = \mathbb{E}\left[\sum_{l=k}^{\tilde{n}-1} R_{t_{l+1}} \middle| X_{t_k} = x\right], \quad V^*(t_k, x) = \sup_{\pi(\cdot)} V(t_k, x; \pi(\cdot)); \tag{10}$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

at any time t_k, where $k = \tilde{n}, \tilde{n}+1, \ldots, n-1$, with the state $x \in \mathcal{X}$, for any policy $\pi(\cdot)$, $V(t_k, x; \pi(\cdot)) = V^*(t_k, x) = 0$. In particular, if $\mathcal{P}(\mathcal{A})$ contains only all Dirac measures over the action space $\mathcal{A}$, which is the case in the DH approach of Bühler *et al.* (2019) (see Appendix A for more details), the value function and the optimal value function reduce to (4) and (6). With this relaxed setting, solving the optimal hedging strategy H^* boils down to finding the optimal policy $\pi^*(\cdot)$.

3.2. Policy approximation and parameterisation

As the hedging problem has the infinite action space $\mathcal{A}$, tabular solution methods for problems of finite state space and finite action space (such as Q-learning), or value function approximation methods for problems of infinite state space and finite action space (such as deep Q-learning) are not suitable. Instead, a *policy gradient method* is employed.

To this end, the policy $\pi(\cdot)$ is approximated and parametrised by the weights θ_{p} in an artificial neural network (ANN); in turn, denote the policy by $\pi(\cdot; \theta_{\mathrm{p}})$. The ANN $\mathcal{N}_{\mathrm{p}}(\cdot; \theta_{\mathrm{p}})$ (to be defined in (11) below) takes a state $x \in \mathcal{X}$ as the input vector and output parameters of a probability measure in $\mathcal{P}(\mathcal{A})$. In the sequel, the set $\mathcal{P}(\mathcal{A})$ contains all Gaussian measures (see, for example, Wang *et al.* 2020 and Wang & Zhou 2020), in which each has a mean c and a variance d^2, which depend on the state input $x \in \mathcal{X}$ and the ANN weights θ_{p}. Therefore, for any state $x \in \mathcal{X}$,

$$\pi(x; \theta_{\mathrm{p}}) = \mu(\cdot; \theta_{\mathrm{p}}) \sim \text{Gaussian}\left(c(x; \theta_{\mathrm{p}}), d^2(x; \theta_{\mathrm{p}})\right),$$

where $\left(c(x; \theta_{\mathrm{p}}), d^2(x; \theta_{\mathrm{p}})\right) = \mathcal{N}_{\mathrm{p}}(x; \theta_{\mathrm{p}})$.

With such approximation and parameterisation, solving the optimal policy π^* further boils down to finding the optimal ANN weights θ_{p}^*. Hence, denote the value function and the optimal value function in (10) by $V(t_k, x; \theta_{\mathrm{p}})$ and $V\left(t_k, x; \theta_{\mathrm{p}}^*\right)$, for any t_k, where $k = 0, 1, \ldots, \tilde{n}-1$, with $x \in \mathcal{X}$. However, the (optimal) value function still depends on the objective probability measure $\mathbb{P}$, the financial market dynamics, and the mortality model, which are unknown to the RL agent. Before formally introducing the policy gradient methods to tackle this issue, we shall first explicitly construct the ANNs for the approximated policy, as well as for an estimate of the value function (to prepare the algorithm of policy gradient method to be reviewed below).

3.3. Network architecture

As alluded above, in this paper, the ANN involves two parts, which are the policy network and the value function network.

3.3.1. Policy network

Let N_{p} be the number of layers for the policy network. For $l = 0, 1, \ldots, N_{\mathrm{p}}$, let $d_{\mathrm{p}}^{(l)}$ be the dimension of the l-th layer, where the 0-th layer is the input layer; the $1, 2, \ldots, (N_{\mathrm{p}}-1)$-th layers are hidden layers; the N_{p}-th layer is the output layer. In particular, $d_{\mathrm{p}}^{(0)} = p$, which is the number of features in the actuarial and financial parts, and $d_{\mathrm{p}}^{(N_{\mathrm{p}})} = 2$, which outputs the mean c and the variance d^2 of the Gaussian measure. The policy network $\mathcal{N}_{\mathrm{p}} : \mathbb{R}^p \to \mathbb{R}^2$ is defined as, for any $x \in \mathbb{R}^p$,

$$\mathcal{N}_{\mathrm{p}}(x) = \left(W_{\mathrm{p}}^{(N_{\mathrm{p}})} \circ \psi \circ W_{\mathrm{p}}^{(N_{\mathrm{p}}-1)} \circ \psi \circ W_{\mathrm{p}}^{(N_{\mathrm{p}}-2)} \circ \cdots \circ \psi \circ W_{\mathrm{p}}^{(1)}\right)(x), \tag{11}$$

where, for $l = 1, 2, \ldots, N_{\mathrm{p}}$, the mapping $W_{\mathrm{p}}^{(l)} : \mathbb{R}^{d_{\mathrm{p}}^{(l-1)}} \to \mathbb{R}^{d_{\mathrm{p}}^{(l)}}$ is affine, and the mapping $\psi : \mathbb{R}^{d_{\mathrm{p}}^{(l)}} \to \mathbb{R}^{d_{\mathrm{p}}^{(l)}}$ is a componentwise activation function. Let θ_{p} be the parameter vector of the policy network and in turn denote the policy network in (11) by $\mathcal{N}_{\mathrm{p}}(x; \theta_{\mathrm{p}})$, for any $x \in \mathbb{R}^p$.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

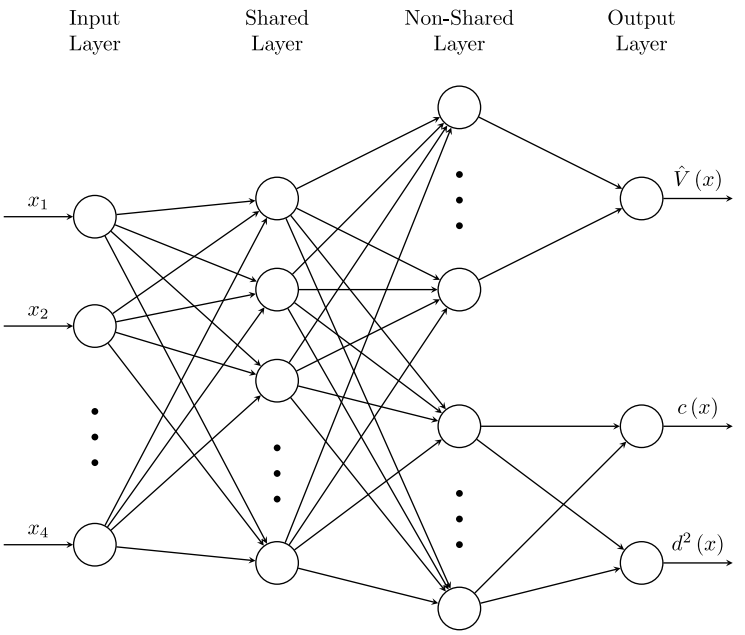


Figure 3. An example of policy and value function artificial neural networks with a shared hidden layer and a non-shared hidden layer.

3.3.2. Value function network

The value function network is constructed similarly as in the policy network, except that all subscripts p (policy) are replaced by v (value). In particular, the value function network $\mathcal{N}_{\mathrm{v}} : \mathbb{R}^p \to \mathbb{R}$ is defined as, for any $x \in \mathbb{R}^p$,

$$\mathcal{N}_{\mathrm{v}}(x) = \left(W_{\mathrm{v}}^{(N_{\mathrm{v}})} \circ \psi \circ W_{\mathrm{v}}^{(N_{\mathrm{v}}-1)} \circ \psi \circ W_{\mathrm{v}}^{(N_{\mathrm{v}}-2)} \circ \cdots \circ \psi \circ W_{\mathrm{v}}^{(1)} \right)(x), \tag{12}$$

which models an approximated value function $\hat{V}$ (see section 3.4 below). Let θ_{v} be the parameter vector of the value function network and in turn denote the value function network in (12) by $\mathcal{N}_{\mathrm{v}}(x; \theta_{\mathrm{v}})$, for any $x \in \mathbb{R}^p$.

3.3.3. Shared layers structure

Since the policy and value function networks should extract features from the input state vector in a similar manner, they are assumed to share the first few layers. More specifically, let $N_{\mathrm{s}} \left(< \min\left\{ N_{\mathrm{p}}, N_{\mathrm{v}} \right\} \right)$ be the number of shared layers for the policy and value function networks; for $l = 1, 2, \ldots, N_{\mathrm{s}}$, $W_{\mathrm{p}}^{(l)} = W_{\mathrm{v}}^{(l)} = W_{\mathrm{s}}^{(l)}$, and hence, for any $x \in \mathbb{R}^p$,

$$\mathcal{N}_{\mathrm{p}}\left(x; \theta_{\mathrm{p}}\right) = \left(W_{\mathrm{p}}^{(N_{\mathrm{p}})} \circ \psi \circ W_{\mathrm{p}}^{(N_{\mathrm{p}}-1)} \circ \cdots \circ \psi \circ W_{\mathrm{p}}^{(N_{\mathrm{s}}+1)} \circ \psi \circ W_{\mathrm{s}}^{(N_{\mathrm{s}})} \circ \cdots \circ \psi \circ W_{\mathrm{s}}^{(1)} \right)(x),$$

$$\mathcal{N}_{\mathrm{v}}(x; \theta_{\mathrm{v}}) = \left(W_{\mathrm{v}}^{(N_{\mathrm{v}})} \circ \psi \circ W_{\mathrm{v}}^{(N_{\mathrm{v}}-1)} \circ \cdots \circ \psi \circ W_{\mathrm{v}}^{(N_{\mathrm{s}}+1)} \circ \psi \circ W_{\mathrm{s}}^{(N_{\mathrm{s}})} \circ \cdots \circ \psi \circ W_{\mathrm{s}}^{(1)} \right)(x).$$

Let θ be the parameter vector of the policy and value function networks. Figure 3 depicts such a shared layers structure.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

3.4. Proximal policy optimisation: a temporal-difference policy gradient method

A policy gradient method entails that, starting from initial ANN weights $\theta^{(0)}$, and via interacting with the MDP environment to observe the states and collect the reward signals, the RL agent gradually updates the ANN weights, by the *(stochastic) gradient ascent* on a certain *surrogate performance measure* defined for the ANN weights. That is, at each update step $u = 1, 2, \ldots,$

$$\theta^{(u)} = \theta^{(u-1)} + \alpha \nabla_\theta \widehat{\mathcal{J}^{(u-1)}\left(\theta^{(u-1)}\right)}, \tag{13}$$

where the hyperparameter $\alpha \in [0, 1]$ is the learning rate of the RL agent, and, based on the experienced episode(s), $\nabla_\theta \widehat{\mathcal{J}^{(u-1)}\left(\theta^{(u-1)}\right)}$ is the estimated gradient of the surrogate performance measure $\mathcal{J}^{(u-1)}(\cdot)$ evaluating at $\theta = \theta^{(u-1)}$.

REINFORCE, which is pioneered by Williams (1992), is a Monte Carlo policy gradient method, which updates the ANN weights by each episode. As this paper applies a temporal-difference (TD) policy gradient method, we relegate the review of REINFORCE to Appendix B, where the *Policy Gradient Theorem*, the foundation of any policy gradient methods, is presented.

PPO, which is pioneered by Schulman *et al.* (2017), is a TD policy gradient method, which updates the ANN weights by a batch of $K \in \mathbb{N}$ realisations. At each update step $u = 1, 2, \ldots,$ based on the ANN weights $\theta^{(u-1)}$, and thus the policy $\pi\left(\cdot; \theta_{\mathrm{p}}^{(u-1)}\right)$, the RL agent experiences $E^{(u)} \in \mathbb{N}$ realised episodes for the K realisations.

- If $E^{(u)} = 1$, the episode is given by

$$\left\{\ldots, x_{t_{K_s^{(u)}}}^{(u-1)}, h_{t_{K_s^{(u)}}}^{(u-1)}, x_{t_{K_s^{(u)}+1}}^{(u-1)}, r_{t_{K_s^{(u)}+1}}^{(u-1)}, h_{t_{K_s^{(u)}+1}}^{(u-1)}, \right.$$
$$\left. \ldots, x_{t_{K_s^{(u)}+K-1}}^{(u-1)}, r_{t_{K_s^{(u)}+K-1}}^{(u-1)}, h_{t_{K_s^{(u)}+K-1}}^{(u-1)}, x_{t_{K_s^{(u)}+K}}^{(u-1)}, r_{t_{K_s^{(u)}+K}}^{(u-1)}, \ldots \right\},$$

  where $K_s^{(u)} = 0, 1, \ldots, \tilde{n} - 1$, such that the time $t_{K_s^{(u)}}$ is when the episode is initiated in this update, and $h_{t_k}^{(u-1)}$, for $k = 0, 1, \ldots, \tilde{n} - 1$, is the time t_k realised hedging strategy being sampled from the Gaussian distribution with the mean $c\left(x_{t_k}^{(u-1)}; \theta_{\mathrm{p}}^{(u-1)}\right)$ and the variance $d^2\left(x_{t_k}^{(u-1)}; \theta_{\mathrm{p}}^{(u-1)}\right)$; necessarily, $\tilde{n} - K_s^{(u)} \geq K$.
- If $E^{(u)} = 2, 3, \ldots,$ the episodes are given by

$$\left\{\ldots, x_{t_{K_s^{(u)}}}^{(u-1,1)}, h_{t_{K_s^{(u)}}}^{(u-1,1)}, x_{t_{K_s^{(u)}+1}}^{(u-1,1)}, r_{t_{K_s^{(u)}+1}}^{(u-1,1)}, h_{t_{K_s^{(u)}+1}}^{(u-1,1)}, \right.$$
$$\left. \ldots, x_{t_{\tilde{n}(1)-1}}^{(u-1,1)}, r_{t_{\tilde{n}(1)-1}}^{(u-1,1)}, h_{t_{\tilde{n}(1)-1}}^{(u-1,1)}, x_{t_{\tilde{n}(1)}}^{(u-1,1)}, r_{t_{\tilde{n}(1)}}^{(u-1,1)} \right\},$$
$$\left\{x_{t_0}^{(u-1,2)}, h_{t_0}^{(u-1,2)}, x_{t_1}^{(u-1,2)}, r_{t_1}^{(u-1,2)}, h_{t_1}^{(u-1,2)}, \right.$$
$$\left. \ldots, x_{t_{\tilde{n}(2)-1}}^{(u-1,2)}, r_{t_{\tilde{n}(2)-1}}^{(u-1,2)}, h_{t_{\tilde{n}(2)-1}}^{(u-1,2)}, x_{t_{\tilde{n}(2)}}^{(u-1,2)}, r_{t_{\tilde{n}(2)}}^{(u-1,2)} \right\},$$
$$\ldots,$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

$$\left\{x_{t_0}^{(u-1,E^{(u)}-1)}, h_{t_0}^{(u-1,E^{(u)}-1)}, x_{t_1}^{(u-1,E^{(u)}-1)}, r_{t_1}^{(u-1,E^{(u)}-1)}, h_{t_1}^{(u-1,E^{(u)}-1)}, \right.$$
$$\left. \ldots, x_{t_{\tilde{n}(E^{(u)}-1)-1}}^{(u-1,E^{(u)}-1)}, r_{t_{\tilde{n}(E^{(u)}-1)-1}}^{(u-1,E^{(u)}-1)}, h_{t_{\tilde{n}(E^{(u)}-1)-1}}^{(u-1,E^{(u)}-1)}, x_{t_{\tilde{n}(E^{(u)}-1)}}^{(u-1,E^{(u)}-1)}, r_{t_{\tilde{n}(E^{(u)}-1)}}^{(u-1,E^{(u)}-1)} \right\},$$
$$\left\{x_{t_0}^{(u-1,E^{(u)})}, h_{t_0}^{(u-1,E^{(u)})}, x_{t_1}^{(u-1,E^{(u)})}, r_{t_1}^{(u-1,E^{(u)})}, h_{t_1}^{(u-1,E^{(u)})}, \right.$$
$$\left. \ldots, x_{t_{K_f^{(u)}-1}}^{(u-1,E^{(u)})}, r_{t_{K_f^{(u)}-1}}^{(u-1,E^{(u)})}, h_{t_{K_f^{(u)}-1}}^{(u-1,E^{(u)})}, x_{t_{K_f^{(u)}}}^{(u-1,E^{(u)})}, r_{t_{K_f^{(u)}}}^{(u-1,E^{(u)})}, \ldots \right\},$$

where $K_f^{(u)} = 1, 2, \ldots, \tilde{n}^{(E^{(u)})}$, such that the time $t_{K_f^{(u)}}$ is when the last episode is finished (but not necessarily terminated) in this update; necessarily, $\tilde{n}^{(1)} - K_s^{(u)} + \sum_{e=2}^{E^{(u)}-1} \tilde{n}^{(e)} + K_f^{(u)} = K$.

The surrogate performance measure of PPO consists of three components. In the following, fix an update step $u = 1, 2, \ldots$.

Inspired by Schulman *et al.* (2015), in which the time-0 value function difference between two policies is shown to be equal to the expected advantage, together with importance sampling and KL divergence constraint reformulation, the first component in the surrogate performance measure of PPO is given by:

- if $E^{(u)} = 1$,

$$L_{\text{CLIP}}^{(u-1)}(\theta_{\text{p}}) = \mathbb{E}\left[\sum_{k=K_s^{(u)}}^{K_s^{(u)}+K-1} \min\left\{q_{t_k}^{(u-1)} \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1)}, \text{clip}\left(q_{t_k}^{(u-1)}, 1-\epsilon, 1+\epsilon\right) \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1)}\right\}\right],$$

where the importance sampling ratio $q_{t_k}^{(u-1)} = \frac{\phi\left(H_{t_k}^{(u-1)}; X_{t_k}^{(u-1)}, \theta_{\text{p}}\right)}{\phi\left(H_{t_k}^{(u-1)}; X_{t_k}^{(u-1)}, \theta_{\text{p}}^{(u-1)}\right)}$, in which $\phi\left(\cdot; X_{t_k}^{(u-1)}, \theta_{\text{p}}\right)$ is the Gaussian density function with mean $c\left(X_{t_k}^{(u-1)}; \theta_{\text{p}}\right)$ and variance $d^2\left(X_{t_k}^{(u-1)}; \theta_{\text{p}}\right)$, the estimated advantage is evaluated at $\theta_{\text{p}} = \theta_{\text{p}}^{(u-1)}$ and bootstrapped through the approximated value function that

$$\hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1)} = \begin{cases} \sum_{l=k}^{K_s^{(u)}+K-1} R_{t_{l+1}}^{(u-1)} + \hat{V}\left(t_{K_s^{(u)}+K}, X_{t_{K_s^{(u)}+K}}^{(u-1)}; \theta_{\text{v}}^{(u-1)}\right) \\ \qquad\qquad - \hat{V}\left(t_k, X_{t_k}^{(u-1)}; \theta_{\text{v}}^{(u-1)}\right) & \text{if } K_s^{(u)} + K < \tilde{n}, \\ \sum_{l=k}^{\tilde{n}-1} R_{t_{l+1}}^{(u-1)} - \hat{V}\left(t_k, X_{t_k}^{(u-1)}; \theta_{\text{v}}^{(u-1)}\right) & \text{if } K_s^{(u)} + K = \tilde{n}, \end{cases}$$

and the function $\text{clip}\left(q_{t_k}^{(u-1)}, 1-\epsilon, 1+\epsilon\right) = \min\left\{\max\left\{q_{t_k}^{(u-1)}, 1-\epsilon\right\}, 1+\epsilon\right\}$. The approximated value function $\hat{V}$ is given by the output of the value network, i.e. $\hat{V}\left(t_k, X_{t_k}^{(u-1)}; \theta_{\text{v}}^{(u-1)}\right) = \mathcal{N}_{\text{v}}\left(X_{t_k}^{(u-1)}; \theta_{\text{v}}^{(u-1)}\right)$ as defined in (12) for $k = 0, 1, \ldots, \tilde{n} - 1$.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

- if $E^{(u)} = 2, 3, \ldots,$

$$L_{\text{CLIP}}^{(u-1)}(\theta_{\text{p}}) = \mathbb{E}\Bigg[\sum_{k=K_s^{(u)}}^{\tilde{n}^{(1)}-1} \min\left\{q_{t_k}^{(u-1,1)} \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,1)}, \text{clip}\left(q_{t_k}^{(u-1,1)}, 1-\epsilon, 1+\epsilon\right) \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,1)}\right\}$$
$$+ \sum_{e=2}^{E^{(u)}-1} \sum_{k=0}^{\tilde{n}^{(e)}-1} \min\left\{q_{t_k}^{(u-1,e)} \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,e)}, \text{clip}\left(q_{t_k}^{(u-1,e)}, 1-\epsilon, 1+\epsilon\right) \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,e)}\right\}$$
$$+ \sum_{k=0}^{K_f^{(u)}-1} \min\left\{q_{t_k}^{(u-1,E^{(u)})} \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,E^{(u)})}, \text{clip}\left(q_{t_k}^{(u-1,E^{(u)})}, 1-\epsilon, 1+\epsilon\right) \hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,E^{(u)})}\right\}\Bigg].$$

Similar to REINFORCE in Appendix B, the second component in the surrogate performance measure of PPO minimises the loss between the bootstrapped sum of reward signals and the approximated value function. To this end, define:

- if $E^{(u)} = 1,$

$$L_{\text{VF}}^{(u-1)}(\theta_{\text{v}}) = \mathbb{E}\left[\sum_{k=K_s^{(u)}}^{K_s^{(u)}+K-1} \left(\hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1)} + \hat{V}\left(t_k, X_{t_k}^{(u-1)}; \theta_{\text{v}}^{(u-1)}\right) - \hat{V}\left(t_k, X_{t_k}^{(u-1)}; \theta_{\text{v}}\right)\right)^2\right];$$

- if $E^{(u)} = 2, 3, \ldots,$

$$L_{\text{VF}}^{(u-1)}(\theta_{\text{v}}) = \mathbb{E}\Bigg[\sum_{k=K_s^{(u)}}^{\tilde{n}^{(1)}-1} \left(\hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,1)} + \hat{V}\left(t_k, X_{t_k}^{(u-1,1)}; \theta_{\text{v}}^{(u-1)}\right) - \hat{V}\left(t_k, X_{t_k}^{(u-1,1)}; \theta_{\text{v}}\right)\right)^2$$
$$+ \sum_{e=2}^{E^{(u)}-1} \sum_{k=0}^{\tilde{n}^{(e)}-1} \left(\hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,e)} + \hat{V}\left(t_k, X_{t_k}^{(u-1,e)}; \theta_{\text{v}}^{(u-1)}\right) - \hat{V}\left(t_k, X_{t_k}^{(u-1,e)}; \theta_{\text{v}}\right)\right)^2$$
$$+ \sum_{k=0}^{K_f^{(u)}-1} \left(\hat{A}_{\theta_{\text{p}}^{(u-1)},t_k}^{(u-1,E^{(u)})} + \hat{V}\left(t_k, X_{t_k}^{(u-1,E^{(u)})}; \theta_{\text{v}}^{(u-1)}\right) - \hat{V}\left(t_k, X_{t_k}^{(u-1,E^{(u)})}; \theta_{\text{v}}\right)\right)^2\Bigg].$$

Finally, to encourage the RL agent exploring the MDP environment, the third component in the surrogate performance measure of PPO is the entropy bonus. Based on the Gaussian density function, define

- if $E^{(u)} = 1,$

$$L_{\text{EN}}^{(u-1)}(\theta_{\text{p}}) = \mathbb{E}\left[\sum_{k=K_s^{(u)}}^{K_s^{(u)}+K-1} \ln d\left(X_{t_k}^{(u-1)}; \theta_{\text{p}}\right)\right];$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

- if $E^{(u)} = 2, 3, \ldots,$

$$L_{\mathrm{EN}}^{(u-1)}(\theta_{\mathrm{p}}) = \mathbb{E}\left[\sum_{k=K_s^{(u)}}^{\tilde{n}^{(1)}-1} \ln d\left(X_{t_k}^{(u-1,1)}; \theta_{\mathrm{p}}\right) + \sum_{e=2}^{E^{(u)}-1} \sum_{k=0}^{\tilde{n}^{(e)}-1} \ln d\left(X_{t_k}^{(u-1,e)}; \theta_{\mathrm{p}}\right) + \sum_{k=0}^{K_f^{(u)}-1} \ln d\left(X_{t_k}^{\left(u-1,E^{(u)}\right)}; \theta_{\mathrm{p}}\right)\right].$$

Therefore, the surrogate performance measure of PPO is given by:

$$\mathcal{J}^{(u-1)}(\theta) = L_{\mathrm{CLIP}}^{(u-1)}(\theta_{\mathrm{p}}) - c_1 L_{\mathrm{VF}}^{(u-1)}(\theta_{\mathrm{v}}) + c_2 L_{\mathrm{EN}}^{(u-1)}(\theta_{\mathrm{p}}), \tag{14}$$

where the hyperparameters $c_1, c_2 \in [0, 1]$ are the loss coefficients of the RL agent. Its estimated gradient, based on the K realisations, is then computed via automatic differentiation; see, for example, Baydin *et al.* (2018).

4. Illustrative Example Revisited: Training Phase

Recall that, in the training phase, the insurer constructs a model of the market environment for an MDP training environment, while the RL agent, which does not know any specifics of this MDP environment, observes states and receives the anchor-hedging reward signals in (9) from it and hence gradually learns the hedging strategy by the PPO algorithm reviewed in the last section. This section revisits the illustrative example in section 2.4 via the two-phase RL approach in the training phase.

4.1. Markov decision process training environment

The model of the market environment is the BS and the CFM in the financial and the actuarial parts. However, unlike the model following the market environment to write a single contract to a single policyholder, for effective training, the insurer writes identical contracts to N homogeneous policyholders in the training environment. Because of the homogeneity of the contracts and the policyholders, for all $i = 1, 2, \ldots, N$, $x_i = x$, $\rho^{(i)} = \rho$, $m^{(i)} = m$, $G_M^{(i)} = G_M$, $G_D^{(i)} = G_D$, $m_e^{(i)} = m_e$, and $F_t^{(i)} = F_t = \rho S_t e^{-mt}$, for $t \in [0, T]$.

At any time $t \in [0, T]$, the future gross liability of the insurer accumulated to the maturity is thus $(G_M - F_T)_+ \sum_{i=1}^N J_T^{(i)} + \sum_{i=1}^N e^{r\left(T - T_x^{(i)}\right)} \left(G_D - F_{T_x^{(i)}}\right)_+ \mathbb{1}_{\{T_x^{(i)} < T\}} J_t^{(i)}$, and its time-$t$ discounted value is

$$V_t^{\mathrm{GL}} = e^{-r(T-t)} \mathbb{E}^{\mathbb{Q}}\left[(G_M - F_T)_+ \sum_{i=1}^N J_T^{(i)} \middle| \mathcal{F}_t\right] + \mathbb{E}^{\mathbb{Q}}\left[\sum_{i=1}^N e^{-r\left(T_x^{(i)} - t\right)} \left(G_D - F_{T_x^{(i)}}\right)_+ \mathbb{1}_{\{T_x^{(i)} < T\}} J_t^{(i)} \middle| \mathcal{F}_t\right]$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

$$= e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\left[(G_M - F_T)_+ |\mathcal{F}_t\right]\sum_{i=1}^{N}\mathbb{E}^{\mathbb{Q}}\left[J_T^{(i)}\middle|\mathcal{F}_t\right]$$

$$+\sum_{i=1}^{N}J_t^{(i)}\mathbb{E}^{\mathbb{Q}}\left[e^{-r\left(T_x^{(i)}-t\right)}\left(G_D - F_{T_x^{(i)}}\right)_+ \mathbb{1}_{\{T_x^{(i)}<T\}}\middle|\mathcal{F}_t\right],$$

where the probability measure $\mathbb{Q}$ defined on $(\Omega, \mathcal{F})$ is an equivalent martingale measure with respect to $\mathbb{P}$. Herein, the probability measure $\mathbb{Q}$ is chosen to be the product measure of each individual equivalent martingale measure in the actuarial or financial part, which implies the independence among the Brownian motion W and the future lifetime $T_x^{(1)}, T_x^{(2)}, \ldots, T_x^{(N)}$, clarifying the first term in the second equality above. The second term in that equality is due to the fact that, for $i = 1, 2, \ldots, N$, the single-jump process $J^{(i)}$ is $\mathbb{F}$-adapted. Under the probability measure $\mathbb{Q}$, all future lifetime are identically distributed and have a CFM $\nu > 0$, which are the same as those under the probability measure $\mathbb{P}$ in section 2.4. Therefore, for any $i = 1, 2, \ldots, N$, and for any $0 \le t \le s \le T$, the conditional survival probability $\mathbb{Q}\left(T_x^{(i)} > s | T_x^{(i)} > t\right) = e^{-\nu(s-t)}$. For each policyholder $i = 1, 2, \ldots, N$, by the independence and the Markov property, for any $0 \le t \le s \le T$,

$$\mathbb{E}^{\mathbb{Q}}\left[J_s^{(i)}\middle|\mathcal{F}_t\right] = \mathbb{E}^{\mathbb{Q}}\left[J_s^{(i)}\middle|J_t^{(i)}\right] = \begin{cases}\mathbb{Q}\left(T_x^{(i)} > s | T_x^{(i)} \le t\right) = 0 & \text{if} \quad T_x^{(i)}(\omega) \le t \\ \mathbb{Q}\left(T_x^{(i)} > s | T_x^{(i)} > t\right) = e^{-\nu(s-t)} & \text{if} \quad T_x^{(i)}(\omega) > t\end{cases}. \tag{15}$$

Moreover, under the probability measure $\mathbb{Q}$, for any $t \in [0, T]$, $dF_t = (r - m)\, F_t dt + \sigma F_t dW_t^{\mathbb{Q}}$, where $W^{\mathbb{Q}} = \left\{W_t^{\mathbb{Q}}\right\}_{t\in[0,T]}$ is the standard Brownian motion under the probability measure $\mathbb{Q}$. Hence, the time-t value of the discounted future gross liability, for $t \in [0, T]$, is given by

$$V_t^{\mathrm{GL}} = e^{-\nu(T-t)}\left(G_M e^{-r(T-t)}\Phi\left(-d_2\left(t, G_M\right)\right) - F_t e^{-m(T-t)}\Phi\left(-d_1\left(t, G_M\right)\right)\right)\sum_{i=1}^{N}J_t^{(i)}$$

$$+\int_t^T\left(G_D e^{-r(T-s)}\Phi\left(-d_2\left(s, G_D\right)\right) - F_t e^{-m(T-s)}\Phi\left(-d_1\left(s, G_D\right)\right)\right)\nu e^{-\nu(s-t)}ds\sum_{i=1}^{N}J_t^{(i)},$$

where, for $s \in [0, T)$ and $G > 0$, $d_1(s, G) = \frac{\ln\left(\frac{F_s}{G}\right) + \left(r - m + \frac{\sigma^2}{2}(T-s)\right)}{\sigma\sqrt{T-s}}$, $d_2(s, G) = d_1(s, G) - \sigma\sqrt{T-s}$, $d_1(T, G) = \lim_{s\to T^-} d_1(s, G)$, $d_2(T, G) = d_1(T, G)$, and $\Phi(\cdot)$ is the standard Gaussian distribution function. Note that $\sum_{i=1}^{N} J_t^{(i)}$ represents the number of surviving policyholders at time $t \in [0, T]$.

As for the cumulative future rider charge to be collected by the insurer from any time $t \in [0, T]$ onward, it is given by $\sum_{i=1}^{N}\int_t^T m_e F_s J_s^{(i)} e^{r(T-s)}ds$, and its time-$t$ discounted value is

$$V_t^{\mathrm{RC}} = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\left[\sum_{i=1}^{N}\int_t^T m_e F_s J_s^{(i)} e^{r(T-s)}ds\middle|\mathcal{F}_t\right] = \sum_{i=1}^{N}\int_t^T m_e e^{-r(s-t)}\mathbb{E}^{\mathbb{Q}}\left[F_s|F_t\right]\mathbb{E}^{\mathbb{Q}}\left[J_s^{(i)}\middle|J_t^{(i)}\right]ds,$$

where the second equality is again due to the independence and the Markov property. Under the probability measure $\mathbb{Q}$, $\mathbb{E}^{\mathbb{Q}}\left[F_s|F_t\right] = e^{(r-m)(s-t)}F_t$. Together with (15),

$$V_t^{\mathrm{RC}} = \frac{1 - e^{-(m+\nu)(T-t)}}{m + \nu} m_e F_t \sum_{i=1}^{N} J_t^{(i)}.$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Therefore, the time-t net liability of the insurer, for $t \in [0, T]$, is given by

$$L_t = V_t^{\mathrm{GL}} - V_t^{\mathrm{RC}} = \left(e^{-\nu(T-t)} \left(G_M e^{-r(T-t)} \Phi\left(-d_2\left(t, G_M\right)\right) - F_t e^{-m(T-t)} \Phi\left(-d_1\left(t, G_M\right)\right) \right) \right.$$
$$+ \int_t^T \left(G_D e^{-r(T-s)} \Phi\left(-d_2\left(s, G_D\right)\right) - F_t e^{-m(T-s)} \Phi\left(-d_1\left(s, G_D\right)\right) \right)$$
$$\left. \nu e^{-\nu(s-t)} ds - \frac{1 - e^{-(m+\nu)(T-t)}}{m+\nu} m_e F_t \right) \sum_{i=1}^{N} J_t^{(i)}, \tag{16}$$

which contributes parts of the reward signals in (9). The time-t value of the insurer's hedging portfolio, for $t \in [0, T]$, as in (1), is given by: $P_0 = 0$, and if $t \in \left(t_k, t_{k+1}\right]$, for some $k = 0, 1, \ldots, n-1$,

$$P_t = \left(P_{t_k} - H_{t_k} S_{t_k}\right) e^{r(t-t_k)} + H_{t_k} S_t + m_e \int_{t_k}^{t} F_s e^{r(t-s)} \sum_{i=1}^{N} J_s^{(i)} ds$$
$$- \sum_{i=1}^{N} e^{r\left(t - T_x^{(i)}\right)} \left(G_D - F_{T_x^{(i)}}\right)_+ \mathbb{1}_{\{t_k < T_x^{(i)} \le t < T\}}, \tag{17}$$

which is also supplied to the reward signals in (9).

At each time t_k, where $k = 0, 1, \ldots, n$, the RL agent is given to observe four features from this MDP environment; these four features are summarised in the state vector

$$X_{t_k} = \left(\ln F_{t_k}, \frac{P_{t_k}}{N}, \frac{\sum_{i=1}^{N} J_{t_k}^{(i)}}{N}, T - t_k \right). \tag{18}$$

The first feature is the natural logarithm of the segregated account value of the policyholder. The second feature is the hedging portfolio value of the insurer, being normalised by the initial number of policyholders. The third feature is the ratio of the number of surviving policyholders with respect to the initial number of policyholders. These features are either log-transformed or normalised to prevent the RL agent from exploring and learning from features with high variability. The last feature is the term to maturity. In particular, when either the third or the last feature first hits zero, i.e. at time $t_{\tilde{n}}$, an episode is terminated. The state space $\mathcal{X} = \mathbb{R} \times \mathbb{R} \times [0, 1/N, 2/N, \ldots, 1] \times \{0, t_1, t_2, \ldots, T\}$.

Recall that, at each time t_k, where $k = 0, 1, \ldots, \tilde{n} - 1$, with the state vector (18) being the input, the output of the policy network in (11) is the mean $c\left(X_{t_k}; \theta_{\mathrm{p}}\right)$ and the variance $d^2\left(X_{t_k}; \theta_{\mathrm{p}}\right)$ of a Gaussian measure; herein, the Gaussian measure represents the distribution of the average number of shares of the risky asset being held by the insurer at the time t_k for each surviving policyholder. Hence, for $k = 0, 1, \ldots, \tilde{n} - 1$, the hedging strategy H_{t_k} in (17) is given by $H_{t_k} = \overline{H}_{t_k} \sum_{i=1}^{N} J_{t_k}^{(i)}$, where $\overline{H}_{t_k}$ is sampled from the Gaussian measure. Since the hedging strategy is assumed to be Markovian with respect to the state vector, it can be shown, albeit tedious, that the state vector, in (18), and the hedging strategy together satisfy the Markov property in (3).

Also recall that the infant RL agent is trained in the MDP environment with multiple homogeneous policyholders. The RL agent should then effectively update the ANN weights θ and learn the hedging strategies, via a more direct inference on the force of mortality from the third feature in the state vector. The RL agent hedges daily, so that the difference between the consecutive discrete hedging time is $\delta t_k = t_{k+1} - t_k = \frac{1}{252}$, for $k = 0, 1, \ldots, n-1$. In this MDP training environment, the parameters of the model are given in Table 3, but with $N = 500$.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 6. Hyperparameters setting of Proximal Policy Optimisation and neural network.

(a) Hyperparameters for proximal policy optimisation			
Grid-searched		Pre-specified	
Hyperparameter	Value	Hyperparameter	Value
Learning rate α	0.07	Coefficient of value function approximation loss c_1	0.25
Batch size K	2048		
Clip factor ϵ	0.18	Coefficient of entropy bonus c_2	0.01
(b) Hyperparameters for neural network			
Hyperparameter			Value(s)
Number of layers in policy network N_{p}			6
Number of layers in value function network N_{v}			6
Number of shared layers N_{s}			3
Dimension of hidden layers in policy network $d_{\mathrm{p}}^{(l)}$			[32, 64, 128, 64, 32]
Dimension of hidden layers in value function network $d_{\mathrm{v}}^{(l)}$			[32, 64, 128, 64, 32]
Activation function $\psi(\cdot)$			ReLU

4.2. Building reinforcement learning agent

After constructing this MDP training environment, the insurer builds the RL agent which implements the PPO, which was reviewed in section 3.4. Table 6(a) summarises all hyperparameters of the implemented PPO, in which three of them are determined via grid search[2], while the remaining two are fixed a priori since they alter the surrogate performance measure itself, and thus should not be based on grid search. Table 6(b) outlines the hyperparameters of the ANN architecture in section 3.3, which are all pre-specified, in which ReLU stands for Rectified Linear Unit; that is, the componentwise activation function is given by, for any $z \in \mathbb{R}$, $\psi(z) = \max\{z, 0\}$.

4.3. Training of reinforcement learning agent

With all these being set up, the insurer assigns the RL agent experiencing this MDP training environment, in order to observe the state, decide, as well as revise, the hedging strategy, and collect the anchor-hedging reward signal based on (9), as much as possible. Let $\mathcal{U} \in \mathbb{N}$ be the number of update steps in the training environment on the ANN weights. Hence, the policy of the experienced RL agent is given by $\pi\left(\cdot; \theta^{(\mathcal{U})}\right) = \pi\left(\cdot; \theta_{\mathrm{p}}^{(\mathcal{U})}\right)$.

Figure 4 depicts the training log of the RL agent in terms of bootstrapped sum of rewards and batch entropy. In particular, Figure 4(a) shows that the value function in (2) reduces to almost zero after around 10^8 training timesteps, which is equivalent to around 48,828 update steps for the ANN weights; within the same number of training timesteps, Figure 4(b) illustrates a gradual depletion on the batch entropy, and hence, the Gaussian measure gently becomes more concentrating around its mean, which implies that the RL agent *progressively diminishes* the degree of *exploration* on the MDP training environment, while *increases* the degree of *exploitation* on the learned ANN weights.

4.4. Baseline hedging performance

In the final step of the training phase, the trained RL agent is assigned to hedge in simulated scenarios from the same MDP training environment, except that $N = 1$ which is in line with hedging

[2]The grid search was performed using the Hardware-Accelerated Learning cluster in the National Center for Supercomputing Applications; see Kindratenko et al. (2020).

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

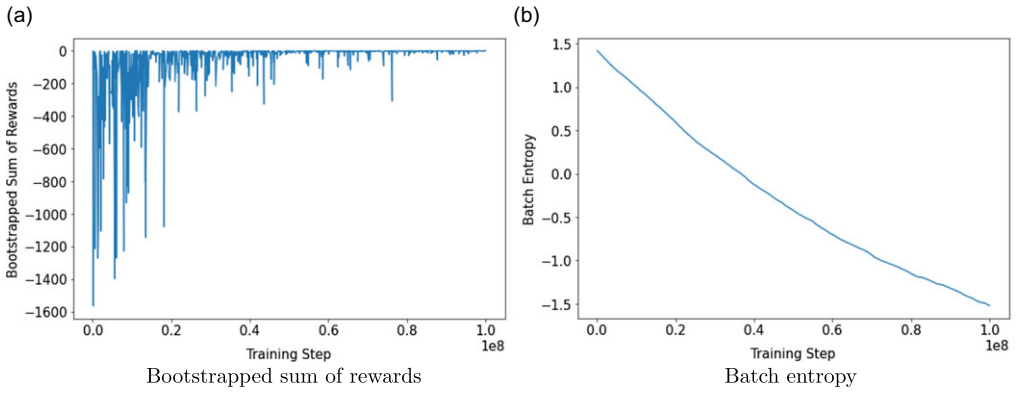


Figure 4. Training log in terms of bootstrapped sum of rewards and batch entropy.

in the market environment. The trained RL agent takes the deterministic action $c\left(\cdot; \theta_{\mathrm{p}}^{(\mathcal{U})}\right)$ which is the mean of the Gaussian measure.

The number of simulated scenarios is 5,000. For each scenario, the insurer documents the realised terminal P&L, i.e. $P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}$. After all scenarios are experienced by the trained RL agent, the insurer examines the baseline hedging performance via the empirical distribution and the summary statistics of the realised terminal P&Ls. The baseline hedging performance of the RL agent is also benchmarked with those by other methods, namely the classical Deltas and the DH; see Appendix C for the implemented hyperparameters of the DH training. The following four classical Deltas are implemented in the simulated scenarios from the training environment, in which the (in)correctness of the Deltas are with respect to the training environment:

- (correct) Delta of the CFM actuarial and BS financial models with the model parameters as in Table 3;
- (incorrect) Delta of the increasing force of mortality (IFM) actuarial and BS financial models, where, for any $i = 1, 2, \ldots, N$, if $T < \overline{b}$, the conditional survival probability $\mathbb{Q}\left(T_x^{(i)} > s | T_x^{(i)} > t\right) = \frac{\overline{b}-s}{\overline{b}-t}$, for any $0 \le t \le s \le T < \overline{b}$, while if $\overline{b} \le T$, the conditional survival probability $\mathbb{Q}\left(T_x^{(i)} > s | T_x^{(i)} > t\right) = \frac{\overline{b}-s}{\overline{b}-t}$, for any $0 \le t \le s < \overline{b} \le T$, and $\mathbb{Q}\left(T_x^{(i)} > s | T_x^{(i)} > t\right) = 0$, for any $0 \le t \le \overline{b} \le s \le T$ or $0 \le \overline{b} \le t \le s \le T$, with the model parameters as in Tables 3(a) and 7;
- (incorrect) Delta in the CFM actuarial and Heston financial models, where, for any $t \in [0, T]$, $dS_t = \mu S_t dt + \sqrt{\Sigma_t} S_t dW_t^{(1)}$, $d\Sigma_t = \kappa\left(\overline{\Sigma} - \Sigma_t\right) dt + \eta\sqrt{\Sigma_t} dW_t^{(2)}$, and $\left\langle W^{(1)}, W^{(2)} \right\rangle_t = \phi t$, with the model parameters as in Tables 3(b) and 8;
- (incorrect) Delta in the IFM actuarial and Heston financial models with the model parameters as in Tables 7 and 8.

Figure 5 shows the empirical density and cumulative distribution functions via the 5,000 realised terminal P&Ls by each hedging approach, while Table 9 outlines the summary statistics of these empirical distributions. To clearly illustrate the comparisons, Figure 6 depicts the empirical density functions via the 5,000 pathwise differences of the realised terminal P&Ls between the RL agent and each of the other approaches, while Table 10 lists the summary statistics of the empirical distributions; for example, comparing with the DH approach, the pathwise difference of the realised terminal P&Ls for the e-th simulated scenario, for $e = 1, 2, \ldots, 5,000$, is calculated by $\left(P_{t_{\tilde{n}}}^{\mathrm{RL}}(\omega_e) - L_{t_{\tilde{n}}}^{\mathrm{RL}}(\omega_e)\right) - \left(P_{t_{\tilde{n}}}^{\mathrm{DH}}(\omega_e) - L_{t_{\tilde{n}}}^{\mathrm{DH}}(\omega_e)\right)$.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 7. Parameters setting of increasing force of mortality actuarial model for Delta.

Parameter	Value
Initial number of policyholder N	1
Initial age of policyholder x	20
Lower bound of uniformly distributed lifetime $\underline{b}$	0
Upper bound of uniformly distributed lifetime $\overline{b}$	50
Investment strategy of policyholders ρ	1.19

Table 8. Parameters setting of Heston financial model for Delta.

Parameter	Value
Risk-free interest rate r	0.02
Risky asset initial price S_0	100
Risky asset drift μ	0.08
Variance initial value Σ_0	0.04
Variance mean reversion rate κ	0.2
Variance long-run average $\overline{\Sigma}$	0.04
Variance volatility η	0.1
Brownian motions correlation ϕ	−0.5

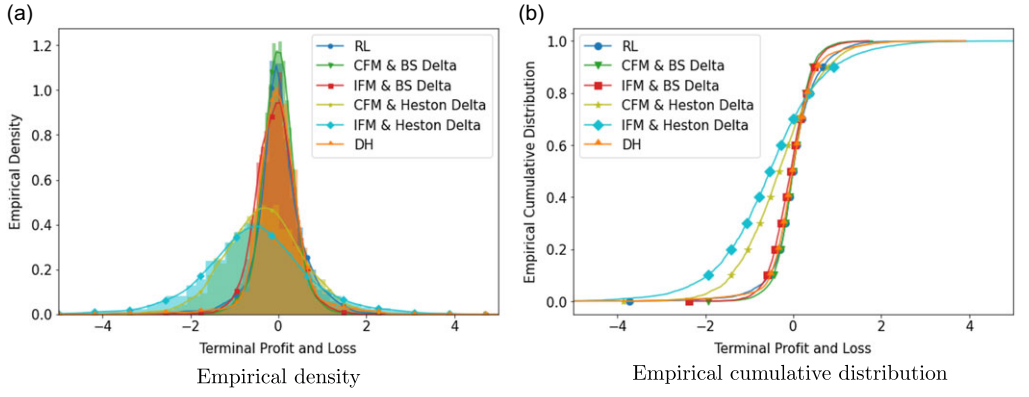


Figure 5. Empirical density and cumulative distribution functions of realised terminal P&Ls by the approaches of reinforcement learning, classical Deltas, and deep hedging.

As expected, the baseline hedging performance of the trained RL agent in this training environment is comparable with those by, the correct CFM and BS Delta, as well as the DH approach. Moreover, the RL agent outperforms all the other three incorrect Deltas, which are based on either incorrect IFM actuarial or Heston financial model, or both.

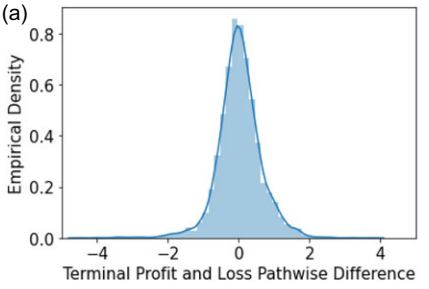
5. Online Learning Phase

Given the satisfactory baseline hedging performance of the experienced RL agent in the MDP training environment, the insurer finally assigns the agent to interact and learn from the market environment.

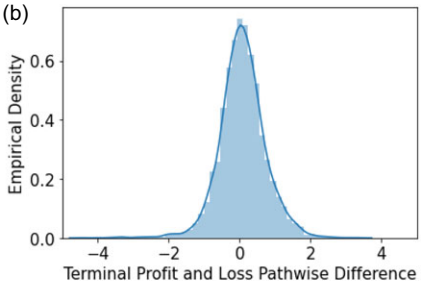
https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 9. Summary statistics of empirical distributions of realised terminal P&Ls by the approaches of reinforcement learning, classical Deltas, and deep hedging.

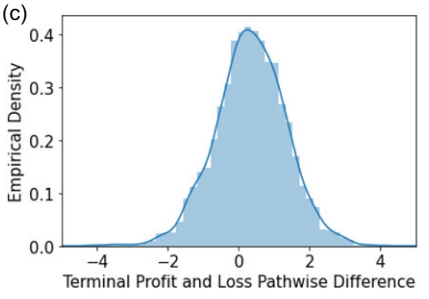
Terminal P&L of hedging approach	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}	$\widehat{\mathrm{RMSE}}$
Reinforcement learning	0.02	−0.01	0.58	−0.54	−0.87	−1.05	−1.43	0.58
CFM & BS Delta	−0.01	0.00	0.38	−0.44	−0.63	−0.70	−0.89	0.38
IFM & BS Delta	−0.06	−0.06	0.45	−0.60	−0.77	−0.85	−1.02	0.45
CFM & Heston Delta	−0.32	−0.33	0.87	−1.41	−1.73	−1.85	−2.17	0.93
IFM & Heston Delta	−0.53	−0.53	1.20	−1.94	−2.48	−2.70	−3.23	1.31
Deep hedging	−0.01	−0.02	0.60	−0.52	−0.71	−1.04	−1.49	0.60



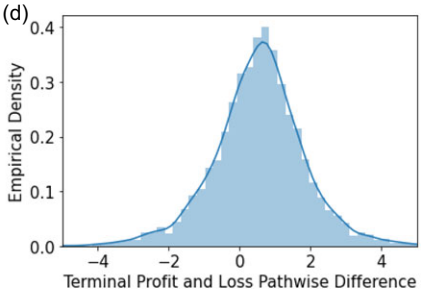
Reinforcement learning versus Delta in constant force of mortality and Black-Scholes models



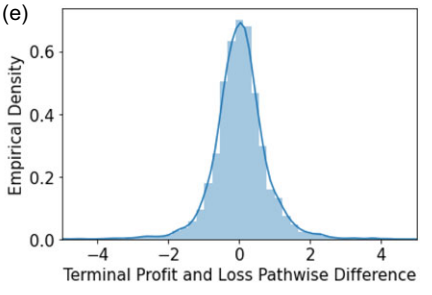
Reinforcement learning versus Delta in increasing force of mortality and Black-Scholes models



Reinforcement learning versus Delta in constant force of mortality and Heston models



Reinforcement learning versus Delta in increasing force of mortality and Heston models



Reinforcement learning versus deep hedging

Figure 6. Empirical density functions of realised pathwise differences of terminal P&Ls comparing with the approaches of classical Deltas and deep hedging.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 10. Summary statistics of empirical distributions of realised pathwise differences of terminal P&Ls comparing with the approaches of classical Deltas and deep hedging.

Pathwise difference of terminal P&Ls comparing with	Mean	Median	Std. Dev.	Probability of non-negativity (%)
CFM & BS delta	0.02	0.01	0.62	50.6
IFM & BS delta	0.08	0.07	0.66	54.7
CFM & Heston delta	0.34	0.34	1.01	64.3
IFM & Heston delta	0.54	0.58	1.29	70.0
Deep hedging	0.02	0.01	0.75	51.3

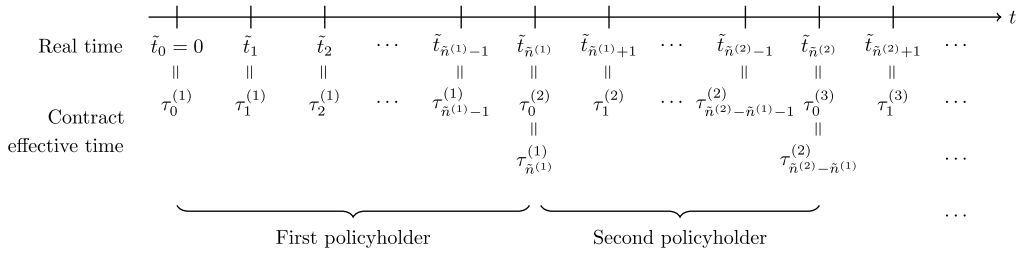


Figure 7. An illustrative timeline with the real time and the contract effective time in the online learning phase.

To distinguish them from the simulated time in the training environment, let $\tilde{t}_k$, for $k = 0, 1, 2, \ldots$, be the real time when the RL agent decides the hedging strategy in the market environment, such that $0 = \tilde{t}_0 < \tilde{t}_1 < \tilde{t}_2 < \cdots$, and $\delta\tilde{t}_k = \tilde{t}_{k+1} - \tilde{t}_k = \frac{1}{252}$. Note that the current time $t = \tilde{t}_0 = 0$ and the RL agent shall hedge daily on behalf of the insurer. At the current time 0, the insurer writes a variable annuity contract with the GMMB and GMDB riders to the first policyholder. When this first contract terminates, due to either the death of the first policyholder or the expiration of the contract, the insurer shall write an identical contract, i.e. contract with the same characteristics, to the second policyholder. And so on. These contract re-establishments ensure that the insurer shall hold only one written variable annuity contract with the GMMB and GMDB riders at a time, and the RL agent shall solely hedge the contract being effective at that moment.

To this end, iteratively, for the ι-th policyholder, where $\iota \in \mathbb{N}$, let $\tilde{t}_{\tilde{n}^{(\iota)}}$ be the first time (right) after the ι-th policyholder dies or the contract expires, for some $\tilde{n}^{(\iota)} = \tilde{n}^{(\iota-1)} + 1, \tilde{n}^{(\iota-1)} + 2, \ldots, \tilde{n}^{(\iota-1)} + n$; that is $\tilde{t}_{\tilde{n}^{(\iota)}} = \min\left\{\tilde{t}_k, k = \tilde{n}^{(\iota-1)} + 1, \tilde{n}^{(\iota-1)} + 2, \ldots, \tilde{n}^{(\iota-1)} + n : \tilde{t}_k - \tilde{t}_{\tilde{n}^{(\iota-1)}} \geq T_{x_\iota}^{(\iota)} \wedge T\right\}$, where, by convention, $\tilde{n}^{(0)} = 0$. Therefore, the contract effective time for the ι-th policyholder $\tau_k^{(\iota)} = \tilde{t}_{\tilde{n}^{(\iota-1)}+k}$, where $\iota \in \mathbb{N}$ and $k = 0, 1, \ldots, \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)}$; in particular, $\tau_0^{(\iota)} = \tilde{t}_{\tilde{n}^{(\iota-1)}}$ is the contract inception time for the ι-th policyholder. Figure 7 depicts one of the possible realisations for clearly illustrating the real time and the contract effective time.

In the online learning phase, the trained RL agent carries on with the PPO of policy gradient methods in the market environment. That is, as in section 3.4, starting from the ANN weights $\theta^{(\mathcal{U})}$ at the current time 0, and via interacting with the market environment to observe the states and collect the reward signals, the RL agent further updates the ANN weights by a batch of $\tilde{K} \in \mathbb{N}$ realisations and the (stochastic) gradient ascent in (13) with the surrogate performance measure in (14), at each update step.

However, there are subtle differences of applying the PPO in the market environment from that in the training environment. At each further update step $\nu = 1, 2, \ldots$, based on the ANN weights $\theta^{(\mathcal{U}+\nu-1)}$, and thus the policy $\pi\left(\cdot; \theta_{\mathrm{p}}^{(\mathcal{U}+\nu-1)}\right)$, the RL agent hedges each effective contract

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

of $\tilde{E}^{(\nu)} \in \mathbb{N}$ realised policyholders for the $\tilde{K} \in \mathbb{N}$ realisations. Indeed, the concept of episodes in the training environment, by the state re-initiation when one episode ends, should be replaced by sequential policyholders in the real-time market environment, via the contract re-establishment when one policyholder dies or contract expires.

- If $\tilde{E}^{(\nu)} = 1$, which is when $(\nu-1)\,\tilde{K}, \nu\tilde{K} \in \left[\tilde{n}^{(\iota-1)}, \tilde{n}^{(\iota)}\right]$, for some $\iota \in \mathbb{N}$, the batch of $\tilde{K}$ realisations is collected solely from the ι-th policyholder. The realisations are given by

$$\left\{\ldots, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s}}, h^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s}}, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+1}}, r^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+1}}, h^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+1}},\right.$$
$$\left.\ldots, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+\tilde{K}-1}}, r^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+\tilde{K}-1}}, h^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+\tilde{K}-1}}, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+\tilde{K}}}, r^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+\tilde{K}}}, \ldots\right\},$$

  where $\tilde{K}^{(\nu)}_s = 0, 1, \ldots, \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)} - 1$, such that the time $\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s}$ is when the first state is observed for the ι-th policyholder in this update; necessarily, $\tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)} - \tilde{K}^{(\nu)}_s \geq \tilde{K}$.
- If $\tilde{E}^{(\nu)} = 2, 3, \ldots$, which is when $(\nu-1)\,\tilde{K} \in \left[\tilde{n}^{(\iota-1)}, \tilde{n}^{(\iota)}\right]$ and $\nu\tilde{K} \in \left[\tilde{n}^{(j-1)}, \tilde{n}^{(j)}\right]$, for some $\iota, j \in \mathbb{N}$ such that $\iota < j$, the batch of $\tilde{K}$ realisations is collected from the ι-th, $(\iota+1)$-th, …, and j-th policyholders; that is, $\tilde{E}^{(\nu)} = j - \iota + 1$. The realisations are given by

$$\left\{\ldots, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s}}, h^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s}}, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+1}}, r^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+1}}, h^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{K}^{(\nu)}_s+1}},\right.$$
$$\left.\ldots, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{n}^{(\iota)}-\tilde{n}^{(i-1)}-1}}, r^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{n}^{(\iota)}-\tilde{n}^{(i-1)}-1}}, h^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{n}^{(\iota)}-\tilde{n}^{(i-1)}-1}}, x^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{n}^{(\iota)}-\tilde{n}^{(i-1)}}}, r^{(\nu-1,\iota)}_{\tau^{(\iota)}_{\tilde{n}^{(\iota)}-\tilde{n}^{(i-1)}}}\right\},$$

$$\left\{x^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_0}, h^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_0}, x^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_1}, r^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_1}, h^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_1},\right.$$
$$\left.\ldots, x^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_{\tilde{n}^{(\iota+1)}-\tilde{n}^{(\iota)}-1}}, r^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_{\tilde{n}^{(\iota+1)}-\tilde{n}^{(\iota)}-1}}, h^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_{\tilde{n}^{(\iota+1)}-\tilde{n}^{(\iota)}-1}}, x^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_{\tilde{n}^{(\iota+1)}-\tilde{n}^{(\iota)}}}, r^{(\nu-1,\iota+1)}_{\tau^{(\iota+1)}_{\tilde{n}^{(\iota+1)}-\tilde{n}^{(\iota)}}}\right\},$$

$$\ldots,$$

$$\left\{x^{(\nu-1,j-1)}_{\tau^{(j-1)}_0}, h^{(\nu-1,j-1)}_{\tau^{(j-1)}_0}, x^{(\nu-1,j-1)}_{\tau^{(j-1)}_1}, r^{(\nu-1,j-1)}_{\tau^{(j-1)}_1}, h^{(\nu-1,j-1)}_{\tau^{(j-1)}_1},\right.$$
$$\left.\ldots, x^{(\nu-1,j-1)}_{\tau^{(j-1)}_{\tilde{n}^{(j-1)}-\tilde{n}^{(j-2)}-1}}, r^{(\nu-1,j-1)}_{\tau^{(j-1)}_{\tilde{n}^{(j-1)}-\tilde{n}^{(j-2)}-1}}, h^{(\nu-1,j-1)}_{\tau^{(j-1)}_{\tilde{n}^{(j-1)}-\tilde{n}^{(j-2)}-1}}, x^{(\nu-1,j-1)}_{\tau^{(j-1)}_{\tilde{n}^{(j-1)}-\tilde{n}^{(j-2)}}}, r^{(\nu-1,j-1)}_{\tau^{(j-1)}_{\tilde{n}^{(j-1)}-\tilde{n}^{(j-2)}}}\right\},$$

$$\left\{x^{(\nu-1,j)}_{\tau^{(j)}_0}, h^{(\nu-1,j)}_{\tau^{(j)}_0}, x^{(\nu-1,j)}_{\tau^{(j)}_1}, r^{(\nu-1,j)}_{\tau^{(j)}_1}, h^{(\nu-1,j)}_{\tau^{(j)}_1},\right.$$
$$\left.\ldots, x^{(\nu-1,j)}_{\tau^{(j)}_{\tilde{K}^{(\nu)}_f-1}}, r^{(\nu-1,j)}_{\tau^{(j)}_{\tilde{K}^{(\nu)}_f-1}}, h^{(\nu-1,j)}_{\tau^{(j)}_{\tilde{K}^{(\nu)}_f-1}}, x^{(\nu-1,j)}_{\tau^{(j)}_{\tilde{K}^{(\nu)}_f}}, r^{(\nu-1,j)}_{\tau^{(j)}_{\tilde{K}^{(\nu)}_f}}, \ldots\right\},$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 11. Hyperparameters setting of Proximal Policy Optimisation for online learning with bolded hyperparameters being different from those for training.

Hyperparameter	Value	Hyperparameter	Value
Learning rate $\tilde{\alpha}$	0.001	Coefficient of value function approximation loss c_1	0.25
Batch size $\tilde{K}$	30		
Clip factor ϵ	0.18	Coefficient of entropy bonus c_2	0.01

where $\tilde{K}_f^{(\nu)} = 1, 2, \ldots, \tilde{n}^{(j)} - \tilde{n}^{(j-1)}$, such that the time $\tau^{(j)}_{\tilde{K}_f^{(\nu)}}$ is when the last state is observed for the j-th policyholder in this update; necessarily, $\tilde{n}^{(j-1)} - \tilde{n}^{(i-1)} + \tilde{K}_f^{(\nu)} - \tilde{K}_s^{(\nu)} = \tilde{K}$.

Moreover, the first two features in the state vector (18) are based on the real-time risky asset price realisation from the market, while all features depend on a particular effective policyholder. For $\iota \in \mathbb{N}$ and $k = 0, 1, \ldots, \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)}$,

$$X^{(\nu-1,\iota)}_{\tau_k^{(\iota)}} = \begin{cases} \left(\ln F^{(\iota)}_{\tau_k^{(\iota)}}, P^{(\iota)}_{\tau_k^{(\iota)}}, 1, T - \left(\tau_k^{(\iota)} - \tau_0^{(\iota)}\right) \right) & \text{if } k = 0, 1, \ldots, \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)} - 1 \\ \left(\ln F^{(\iota)}_{\tau_k^{(\iota)}}, P^{(\iota)}_{\tau_k^{(\iota)}}, 0, T - \left(\tau_k^{(\iota)} - \tau_0^{(\iota)}\right) \right) & \text{if } k = \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)} \text{ and } T^{(\iota)}_{x_\iota} \le T \\ \left(\ln F^{(\iota)}_{\tau_k^{(\iota)}}, P^{(\iota)}_{\tau_k^{(\iota)}}, 1, 0 \right) & \text{if } k = \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)} \text{ and } T^{(\iota)}_{x_\iota} > T \end{cases}, \tag{19}$$

where $F_t^{(\iota)} = \rho^{(\iota)} S_t e^{-m^{(\iota)}\left(t - \tau_0^{(\iota)}\right)}$, if $t \in \left[\tau_0^{(\iota)}, \tilde{t}_{\tilde{n}^{(\iota)}}\right]$, $P^{(\iota)}_{\tau_0^{(\iota)}} = 0$, and

$$\begin{aligned} P^{(\iota)}_{\tau_k^{(\iota)}} = & \left(P^{(\iota)}_{\tau_{k-1}^{(\iota)}} - H^{(\iota)}_{\tau_{k-1}^{(\iota)}} S_{\tau_{k-1}^{(\iota)}} \right) e^{r\left(\tau_k^{(\iota)} - \tau_{k-1}^{(\iota)}\right)} + H^{(\iota)}_{\tau_{k-1}^{(\iota)}} S_{\tau_k^{(\iota)}} + m_e^{(\iota)} \int_{\tau_{k-1}^{(\iota)}}^{\tau_k^{(\iota)}} F_s^{(\iota)} e^{r\left(\tau_k^{(\iota)} - s\right)} J_s^{(\iota)} ds \\ & - \left(G_D - F^{(\iota)}_{T^{(\iota)}_{x_\iota}} \right)_+ \mathbb{1}_{\{\tau_{k-1}^{(\iota)} < T^{(\iota)}_{x_\iota} \le \tau_k^{(\iota)}\}} e^{r\left(\tau_k^{(\iota)} - T^{(\iota)}_{x_\iota}\right)}, \end{aligned}$$

for $k = 1, 2, \ldots, \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)}$. Recall also that the reward signals collecting from the market environment should be based on that in (8); that is, for $\iota \in \mathbb{N}$ and $k = 0, 1, \ldots, \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)}$,

$$R^{(\nu-1,\iota)}_{\tau_k^{(\iota)}} = \begin{cases} 0 & \text{if} \quad k = 0, 1, \ldots, \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)} - 1 \\ -\left(P^{(\iota)}_{\tilde{t}_{\tilde{n}^{(\iota)}}} - L^{(\iota)}_{\tilde{t}_{\tilde{n}^{(\iota)}}} \right)^2 & \text{if} \quad k = \tilde{n}^{(\iota)} - \tilde{n}^{(\iota-1)} \end{cases},$$

in which $L^{(\iota)}_{\tilde{t}_{\tilde{n}^{(\iota)}}} = 0$ if $T^{(\iota)}_{x_\iota} \le T$, and $L^{(\iota)}_{\tilde{t}_{\tilde{n}^{(\iota)}}} = \left(G_M - F^{(\iota)}_{\tau_0^{(\iota)}+T} \right)_+$ if $T^{(\iota)}_{x_\iota} > T$.

Table 11 summarises all hyperparameters of the implemented PPO in the market environment, while the hyperparameters of the ANN architecture are still given in Table 6(b). In the online learning phase, the insurer should choose a smaller batch size $\tilde{K}$ comparing to that in the training phase; this yields a higher updating frequency by the PPO to ensure that the experienced RL agent could revise the hedging strategy within a reasonable amount of time. However, fewer realisations in the batch cause less credible updates; hence, the insurer should also tune down the learning rate $\tilde{\alpha}$, from that in the training phase, to reduce the reliance on each further update step.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

6. Illustrative Example Revisited: Online Learning Phase

This section revisits the illustrative example in section 2.4 via the two-phase RL approach in the online learning phase. In the market environment, the policyholders being sequentially written of the contracts with both GMMB and GMDB riders are homogeneous. Due to contract re-establishments to these sequential homogeneous policyholders, the number and age of policyholders shall be reset to the values as in Table 3(b) at each contract inception time. Furthermore, via the approach discussed in section 2.1.3, to determine the fee structures of each contract at its inception time, the insurer relies on the parameters of the model of the market environment in Table 3, except that now the risky asset initial price therein is replaced by the risky asset price observed at the contract inception time. Note that the fee structures of the first contract are still given as in Table 4, since the risky asset price observed at $t = 0$ is exactly the risky asset initial price.

Let $\mathcal{V} \in \mathbb{N}$ be the number of further update steps in the market environment on the ANN weights. In order to showcase the result that (RLw/OL), the further trained RL agent with the online learning phase, could gradually revise the hedging strategy, from the nearly optimal one in the training environment, to the one in the market environment, we evaluate the hedging performance of RLw/OL on a rolling basis. That is, right after each further update step $v = 1, 2, \ldots, \mathcal{V}$, we first simulate $\tilde{M} = 500$ market scenarios stemming from the real-time realised state vector $x^{(v-1,j)}_{\tau^{(j)}_{\tilde{K}_f^{(v)}}}$ and by implementing the hedging strategy from the updated policy $\pi\left(\cdot; \theta_{\mathrm{p}}^{(\mathcal{U}+v)}\right)$, i.e. the further trained RL agent takes the deterministic action $c\left(\cdot; \theta_{\mathrm{p}}^{(\mathcal{U}+v)}\right)$ which is the mean of the Gaussian measure; we then document the realised terminal P&L, for each of the 500 simulated scenarios, i.e. $P_t^{\mathrm{RLw/OL}}(\omega_e) - L_t(\omega_e)$, for $e = 1, 2, \ldots, 500$, where $t = \tilde{t}_{\tilde{n}(j)}(\omega_e)$ if $\tau^{(j)}_{\tilde{K}_f^{(v)}} < \tilde{t}_{\tilde{n}(j)}$, and $t = \tilde{t}_{\tilde{n}(j+1)}(\omega_e)$ if $\tau^{(j)}_{\tilde{K}_f^{(v)}} = \tilde{t}_{\tilde{n}(j)}$.

Since the state vector $x^{(v-1,j)}_{\tau^{(j)}_{\tilde{K}_f^{(v)}}}$ is realised in real time, the realised terminal P&L in fact depends on, not only the simulated scenarios after each update but also the actual realisation in the market environment. To this end, from the current time 0, we simulate $M = 1,000$ future trajectories in the market environment; for each future trajectory $f = 1, 2, \ldots, 1,000$, the aforementioned realised terminal P&Ls are obtained as $P_t^{\mathrm{RLw/OL}}(\omega_f, \omega_e) - L_t(\omega_f, \omega_e)$, for $e = 1, 2, \ldots, 500$, where $t = \tilde{t}_{\tilde{n}(j)}(\omega_f, \omega_e)$ if $\tau^{(j)}_{\tilde{K}_f^{(v)}}(\omega_f) < \tilde{t}_{\tilde{n}(j)}(\omega_f)$, and $t = \tilde{t}_{\tilde{n}(j+1)}(\omega_f, \omega_e)$ if $\tau^{(j)}_{\tilde{K}_f^{(v)}}(\omega_f) = \tilde{t}_{\tilde{n}(j)}(\omega_f)$.

The rolling-basis hedging performance of RLw/OL is benchmarked with those by, (RLw/oOL) the trained RL agent without the online learning phase, (CD) the correct Delta based on the market environment, and (ID) the incorrect Delta based on the training environment. For the same set of future trajectories ω_f, for $f = 1, 2, \ldots, 1,000$, and the same sets of simulated scenarios ω_e, for $e = 1, 2, \ldots, 500$, the realised terminal P&Ls are also obtained, by implementing each of these benchmark strategies starting from the current time 0, which does not need to be updated throughout; denote the realised terminal P&L as $P_t^{\mathcal{S}}(\omega_f, \omega_e) - L_t(\omega_f, \omega_e)$, where $\mathcal{S} =$ RLw/OL, RLw/oOL, CD, or ID.

This example considers $\mathcal{V} = 25$ further update steps of RLw/OL, for each future trajectory ω_f, where $f = 1, 2, \ldots, 1,000$; as the batch size in the online learning phase $\tilde{K} = 30$, this is equivalent to 750 trading days, which is just less than 3 years (assuming that non-trading days are uniformly spread across a year). For each $f = 1, 2, \ldots, 1,000$, and $v = 1, 2, \ldots, 25$, let $\mu_{\mathcal{S}}^{(v,j)}(\omega_f)$ be the expected terminal P&L, right after the v-th further update step implementing the hedging

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

strategy $\mathcal{S}$ for the future trajectory ω_f:

$$\mu_{\mathcal{S}}^{(v,j)}(\omega_f) = \mathbb{E}\left[P_t^{\mathcal{S}}(\omega_f, \cdot) - L_t(\omega_f, \cdot) \,\middle|\, X_{\tau_{\tilde{K}_f^{(v)}}^{(j)}}^{(v-1,j)} = X_{\tau_{\tilde{K}_f^{(v)}}^{(j)}}^{(v-1,j)}(\omega_f) \right],$$

which is a conditional expectation taking with respect to the scenarios from the time $\tau_{\tilde{K}_f^{(v)}}^{(j)}$ forward; let $\hat{\mu}_{\mathcal{S}}^{(v,j)}(\omega_f)$ be the sample mean of the terminal P&L based on the simulated scenarios:

$$\hat{\mu}_{\mathcal{S}}^{(v,j)}(\omega_f) = \frac{1}{500} \sum_{e=1}^{500} \left(P_t^{\mathcal{S}}(\omega_f, \omega_e) - L_t(\omega_f, \omega_e) \right). \tag{20}$$

Figure 8 plots the sample means of the terminal P&L in (20), right after each further update step and implementing each hedging strategy, in two future trajectories. Firstly, notice that, in both future trajectories, the average hedging performance of RLw/oOL is even worse than that of ID. Secondly, the average hedging performances of RLw/OL between the two future trajectories are substantially different. In the best-case future trajectory, the RLw/OL is able to swiftly self-revise the hedging strategy and hence quickly catch up the average hedging performance of ID by simply twelve further updates on the ANN weights, as well as that of CD in around two years; however, in the *worst-case* future trajectory, within 3 years, the RLw/OL is not able to improve the average hedging performance to even the level of ID, let alone to that of CD.

In view of the second observation above, the hedging performance of RLw/OL should not be concluded for each future trajectory alone; instead, it should be studied among the future trajectories. To this end, for each $f = 1, 2, \ldots, 1,000$, define

$$\nu_{\mathrm{CD}}(\omega_f) = \min\left\{ v = 1, 2, \ldots, 25 : \hat{\mu}_{\mathrm{RLw/OL}}^{(v,j)}(\omega_f) > \hat{\mu}_{\mathrm{CD}}^{(v,j)}(\omega_f) \right\}$$

as the first further update step such that the sample mean of the terminal P&L by RLw/OL is strictly greater than that by CD, for the future trajectory ω_f; herein, let $\min \emptyset = 26$ and also define $t_{\mathrm{CD}}(\omega_f) = \nu_{\mathrm{CD}}(\omega_f) \times \frac{\tilde{K}}{252}$ as the corresponding number of years. Therefore, the estimated proportion of the future trajectories, where RLw/OL is able to exceed the average hedging performance of CD within 3 years, is given by

$$\frac{1}{1,000} \sum_{f=1}^{1,000} \mathbb{1}_{\{t_{\mathrm{CD}}(\omega_f) \le 3\}} = 95.4\%.$$

For each $f = 1, 2, \ldots, 1,000$, define $\nu_{\mathrm{ID}}(\omega_f)$ and $t_{\mathrm{ID}}(\omega_f)$ similarly for comparing RLw/OL with ID. Figure 9 shows the empirical conditional density functions of t_{CD} and t_{ID}, both subject to that RLw/OL exceeds the average hedging performance of CD within 3 years. Table 12 lists the summary statistics of the empirical conditional distributions.

The above analysis obviously neglected the variance, due to the simulated scenarios, of hedging performance by each hedging strategy. In the following, for each future trajectory, we define a refined first further update step such that the expected terminal P&L by RLw/OL is statistically significant to be strictly greater than that by CD. To this end, for each $f = 1, 2, \ldots, 1,000$, and $v = 1, 2, \ldots, 25$, consider the following null and alternative hypotheses:

$$H_{0,\mathcal{S}}^{(v,j)}(\omega_f) : \mu_{\mathrm{RLw/OL}}^{(v,j)}(\omega_f) \le \mu_{\mathcal{S}}^{(v,j)}(\omega_f) \quad \text{versus} \quad H_{1,\mathcal{S}}^{(v,j)}(\omega_f) : \mu_{\mathrm{RLw/OL}}^{(v,j)}(\omega_f) > \mu_{\mathcal{S}}^{(v,j)}(\omega_f),$$

where $\mathcal{S} =$ CD or ID; the analysis before supports this choice of the alternative hypothesis. Define respectively the test statistics and the p-value by

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

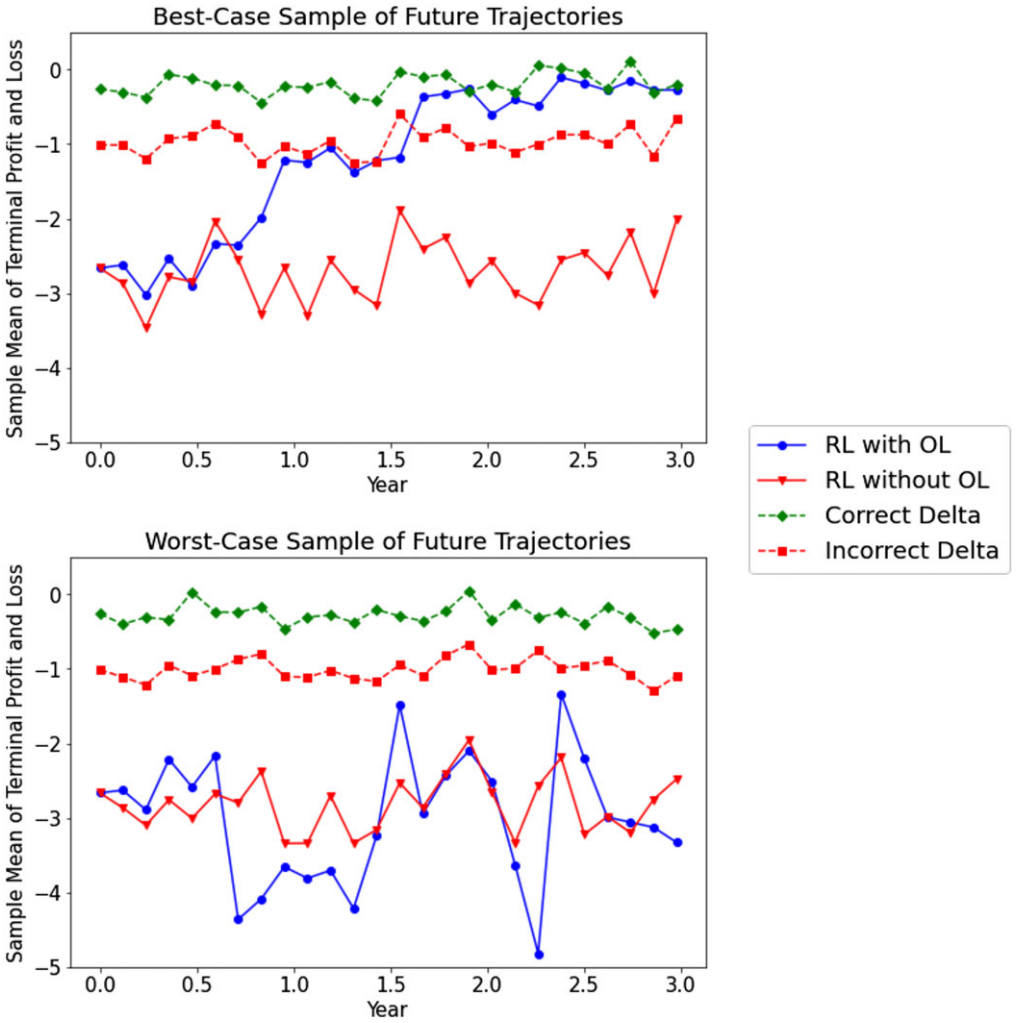


Figure 8. Best-case and worst-case samples of future trajectories for rolling-basis evaluation of reinforcement learning agent with online learning phase, and comparisons with classical Deltas and reinforcement learning agent without online learning phase.

$$\mathcal{T}_{\mathcal{S}}^{(v,j)}(\omega_f) = \frac{\hat{\mu}_{\mathrm{RLw/OL}}^{(v,j)}(\omega_f) - \hat{\mu}_{\mathcal{S}}^{(v,j)}(\omega_f)}{\sqrt{\frac{\hat{\sigma}_{\mathrm{RLw/OL}}^{(v,j)}(\omega_f)^2}{500} + \frac{\hat{\sigma}_{\mathcal{S}}^{(v,j)}(\omega_f)^2}{500}}} \quad \text{and} \quad p_{\mathcal{S}}^{(v,j)}(\omega_f) = \mathbb{P}\Big(T_{\mathcal{S}}(\omega_f) > \mathcal{T}_{\mathcal{S}}^{(v,j)}(\omega_f)\Big),$$

where the random variable $T_{\mathcal{S}}(\omega_f)$ follows a Student's t-distribution with the degree of freedom

$$\mathrm{df}_{\mathcal{S}}^{(v,j)}(\omega_f) = \frac{\left(\frac{\hat{\sigma}_{\mathrm{RLw/OL}}^{(v,j)}(\omega_f)^2}{500} + \frac{\hat{\sigma}_{\mathcal{S}}^{(v,j)}(\omega_f)^2}{500}\right)^2}{\frac{\left(\hat{\sigma}_{\mathrm{RLw/OL}}^{(v,j)}(\omega_f)^2/500\right)^2}{500-1} + \frac{\left(\hat{\sigma}_{\mathcal{S}}^{(v,j)}(\omega_f)^2/500\right)^2}{500-1}},$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

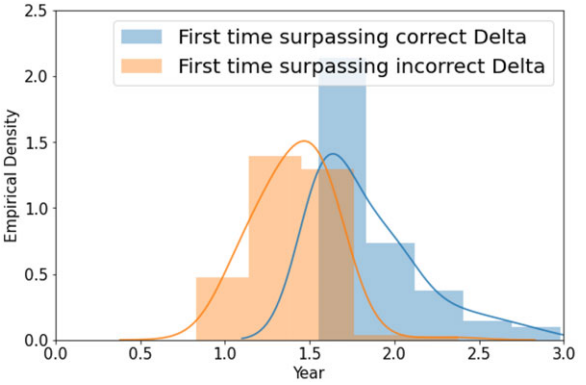


Figure 9. Empirical conditional density functions of first surpassing times conditioning on reinforcement learning agent with online learning phase exceeding correct Delta in terms of sample means of terminal P&L within 3 years.

and the sample variance $\hat{\sigma}_{\mathcal{S}}^{(v,j)}(\omega_f)^2$ of the terminal P&L based on the simulated scenarios is given by

$$\hat{\sigma}_{\mathcal{S}}^{(v,j)}(\omega_f)^2 = \frac{1}{499}\sum_{e=1}^{500}\left(\left(P_t^{\mathcal{S}}(\omega_f,\omega_e) - L_t(\omega_f,\omega_e)\right) - \hat{\mu}_{\mathcal{S}}^{(v,j)}(\omega_f)\right)^2.$$

For a fixed level of significance $\alpha^* \in (0,1)$, if $p_{\mathcal{S}}^{(v,j)}(\omega_f) < \alpha^*$, then the expected terminal P&L by RLw/OL is statistically significant to be strictly greater than that by $\mathcal{S} =$ CD or ID.

In turn, for each $f = 1, 2, \ldots, 1,000$, and for any $\alpha^* \in (0,1)$, define

$$v_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha^*) = \min\left\{v = 1,2,\ldots,25 : p_{\mathcal{S}}^{(v,j)}(\omega_f) < \alpha^*\right\}$$

as the first further update step such that the expected terminal P&L by RLw/OL is statistically significant to be strictly greater than that by $\mathcal{S} =$ CD or ID, for the future trajectory ω_f at the level of significance α^*; again, herein, let $\min\emptyset = 26$ and define $t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha^*) = v_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha^*) \times \frac{\tilde{K}}{252}$ as the corresponding number of years. Table 13 lists the estimated proportion of the future trajectories, where RLw/OL is statistically significant to be able to exceed the expected terminal P&L of $\mathcal{S}$ within 3 years, which is given by $\sum_{f=1}^{1,000} \mathbb{1}_{\{t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha^*)\le 3\}}/1,000$, with various levels of significance.

When the level of significance α^* gradually decreases from 0.20 to 0.01, both estimated proportions, of the future trajectories for RLw/OL being statistically significant to be exceeding CD or ID within 3 years, decline. This is because, for any $\alpha_1^*, \alpha_2^* \in (0,1)$ with $\alpha_1^* \le \alpha_2^*$, and for any ω_f, for $f = 1, 2, \ldots, 1,000$, $t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*) \le 3$ implies that $t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_2^*) \le 3$, and thus, $\mathbb{1}_{\{t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*)\le 3\}} \le \mathbb{1}_{\{t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_2^*)\le 3\}}$, which leads to that $\sum_{f=1}^{1,000} \mathbb{1}_{\{t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*)\le 3\}}/1,000 \le \sum_{f=1}^{1,000} \mathbb{1}_{\{t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_2^*)\le 3\}}/1,000$; indeed, since $t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*) \le 3$, or equivalently $v_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*) \le 25$, we have $p_{\mathcal{S}}^{(v_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*),j)}(\omega_f) < \alpha_1^* \le \alpha_2^*$, and thus

$$v_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_2^*) = \min\left\{v = 1,2,\ldots,25 : p_{\mathcal{S}}^{(v,j)}(\omega_f) < \alpha_2^*\right\} \le v_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*) \le 25,$$

or equivalently $t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_2^*) \le t_{\mathcal{S}}^{\mathrm{p}}(\omega_f;\alpha_1^*) \le 3$. However, notably, the declining rate of the estimated proportion for exceeding CD is greater than that for exceeding ID.

Similar to Figure 9 and Table 12, one can depict the empirical conditional density functions and list the summary statistics of $t_{\mathrm{CD}}^{\mathrm{p}}(\cdot;\alpha^*)$ and $t_{\mathrm{ID}}^{\mathrm{p}}(\cdot;\alpha^*)$, for each level of significance α^*, subject to

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 12. Summary statistics of empirical conditional distributions of first surpassing times conditioning on reinforcement learning agent with online learning phase exceeding correct Delta in terms of sample means of terminal P&L within 3 years.

Reinforcement learning agent with online learning phase first surpassing time to	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
Correct Delta	1.84	1.79	0.32	2.38	2.50	2.66	2.73
Incorrect Delta	1.41	1.43	0.22	1.67	1.67	2.05	2.05

Table 13. Estimated proportions of future trajectories where reinforcement learning agent with online learning phase is statistically significant to be exceeding correct Delta and incorrect Delta within 3 years with various levels of significance.

Estimated proportion of exceeding	$\alpha^* = 0.20$	$\alpha^* = 0.15$	$\alpha^* = 0.10$	$\alpha^* = 0.05$	$\alpha^* = 0.01$
Correct Delta	55.7%	52.1%	47.6%	35.9%	21.8%
Incorrect Delta	96.9%	95.1%	85.0%	70.6%	64.6%

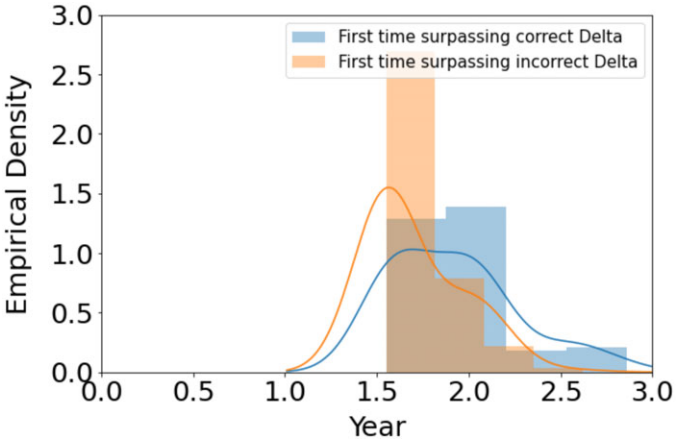


Figure 10. Empirical conditional density functions of first statistically significant surpassing times conditioning on reinforcement learning agent with online learning phase being statistically significant to be exceeding correct Delta within 3 years for 0.1 level of significance.

that RLw/OL is statistically significant to be exceeding CD within 3 years. For example, with $\alpha^* = 0.1$, Figure 10 and Table 14 illustrate that, comparing with Figure 9 and Table 12, the distributions are right-shifted as well as more spread, and the summary statistics are all increased.

Finally, to further examine the hedging performance of RLw/OL in terms of the sample mean of the terminal P&L in (20), as well as take the random future trajectories into account, Figure 11 shows the snapshots of the empirical density functions, among the future trajectories, of the sample mean by each hedging strategy over time at $t = 0, 0.6, 1.2, 1.8, 2.4,$ and 3; Table 15 outlines their summary statistics. Note that, at the current time $t = 0$, since none of the future trajectories has been realised yet, the empirical density functions are given by Dirac delta at the corresponding sample mean by each hedging strategy, which only depends on the simulated scenarios. As the time progresses, one can observe that the empirical density function by RLw/OL is gradually shifting to the right, substantially passing the one by ID and almost catching up the one by CD at $t = 1.8$. This sheds light on the high probability that RLw/OL is able to self-revise the hedging strategy from a very sub-optimal one to a nearly optimal one close to the CD.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 14. Summary statistics of empirical conditional distributions of first statistically significant surpassing times conditioning on reinforcement learning agent with online learning phase being statistically significant to be exceeding correct Delta within 3 years for 0.1 level of significance.

Reinforcement learning agent with online learning phase first surpassing time to	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
Correct Delta	1.92	1.90	0.34	2.50	2.62	2.61	2.70
Incorrect Delta	1.70	1.55	0.24	2.02	2.14	2.07	2.20

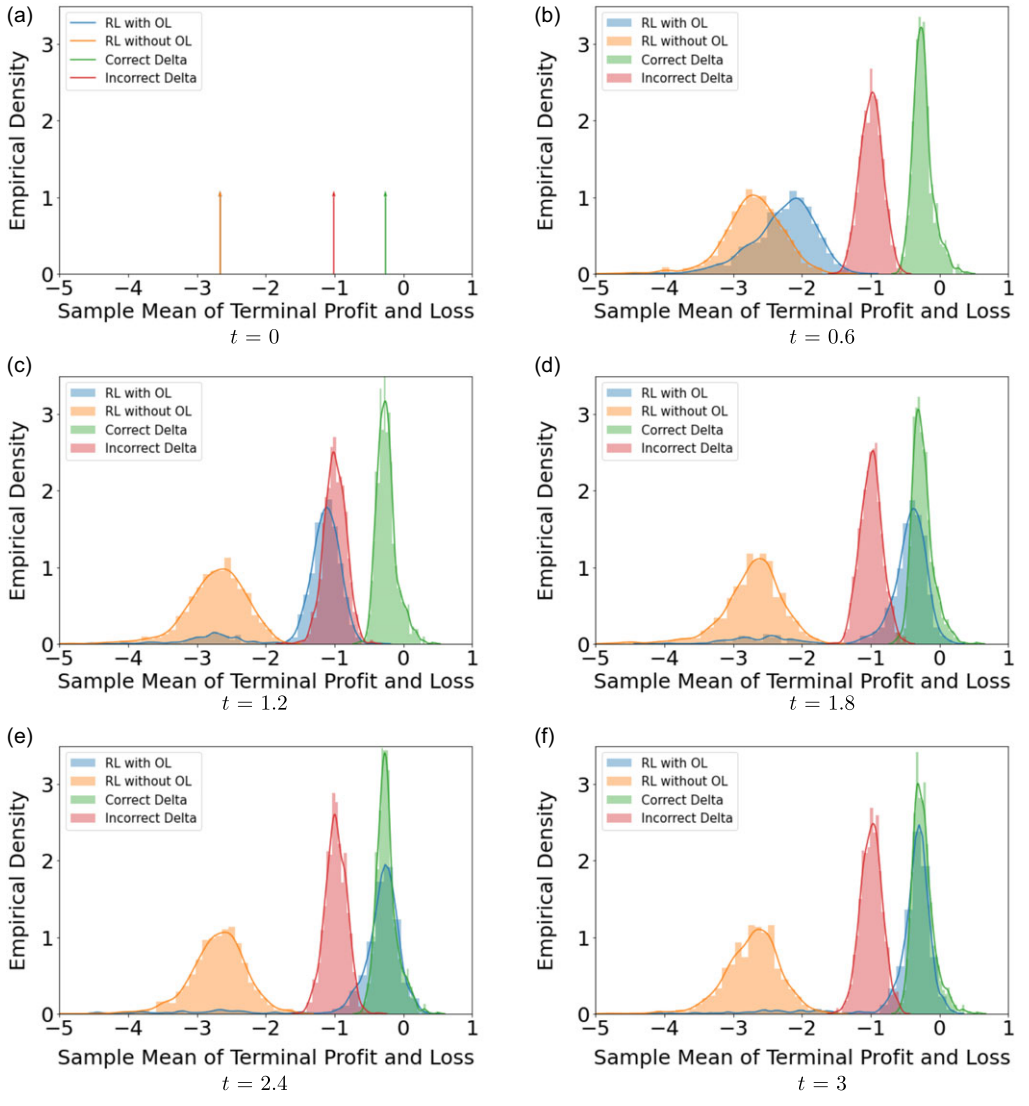


Figure 11. Snapshots of empirical density functions of sample mean of terminal P&L by reinforcement learning agent with online learning phase, reinforcement learning agent without online learning phase, correct Delta, and incorrect Delta at different time points.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Table 15. Summary statistics of empirical distributions of sample mean of terminal P&L by reinforcement learning agent with online learning phase, reinforcement learning agent without online learning phase, correct Delta, and incorrect Delta at different time points.

Sample mean of terminal P&L by	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
RL with OL	−2.66	−2.66	0	−2.66	−2.66	−2.66	−2.66
RL without OL	−2.66	−2.66	0	−2.66	−2.66	−2.66	−2.66
Correct Delta	−0.26	−0.26	0	−0.26	−0.26	−0.26	−0.26
Incorrect Delta	−1.01	−1.01	0	−1.01	−1.01	−1.01	−1.01
(a) $t = 0$							
Sample mean of terminal P&L by	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
RL with OL	−2.27	−2.19	0.45	−2.86	−3.11	−3.17	−3.38
RL without OL	−2.71	−2.69	0.42	−3.20	−3.39	−3.52	−3.76
Correct Delta	−0.24	−0.26	0.15	−0.40	−0.45	−0.46	−0.49
Incorrect Delta	−0.99	−0.99	0.16	−1.20	−1.26	−1.27	−1.31
(b) $t = 0.6$							
Sample mean of terminal P&L by	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
RL with OL	−1.29	−1.14	0.55	−1.83	−2.75	−2.80	−3.10
RL without OL	−2.71	−2.68	0.42	−3.22	−3.45	−3.54	−3.78
Correct Delta	−0.24	−0.26	0.14	−0.41	−0.44	−0.45	−0.48
Incorrect Delta	−0.99	−0.99	0.16	−1.19	−1.25	−1.27	−1.33
(c) $t = 1.2$							
Sample mean of terminal P&L by	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
RL with OL	−0.63	−0.43	0.69	−1.14	−2.50	−2.52	−2.94
RL without OL	−2.70	−2.67	0.43	−3.22	−3.42	−3.58	−3.86
Correct Delta	−0.25	−0.27	0.15	−0.41	−0.45	−0.47	−0.50
Incorrect Delta	−0.99	−0.99	0.15	−1.20	−1.25	−1.27	−1.31
(d) $t = 1.8$							
Sample mean of terminal P&L by	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
RL with OL	−0.46	−0.33	0.71	−0.72	−2.24	−2.13	−3.24
RL without OL	−2.69	−2.66	0.41	−3.18	−3.40	−3.48	−3.69
Correct Delta	−0.24	−0.26	0.15	−0.40	−0.45	−0.46	−0.50
Incorrect Delta	−0.98	−0.98	0.15	−1.18	−1.24	−1.26	−1.30
(e) $t = 2.4$							
Sample mean of terminal P&L by	Mean	Median	Std. Dev.	VaR_{90}	VaR_{95}	TVaR_{90}	TVaR_{95}
RL with OL	−0.45	−0.33	0.56	−0.66	−1.66	−1.75	−2.59
RL without OL	−2.71	−2.68	0.41	−3.24	−3.38	−3.49	−3.68
Correct Delta	−0.24	−0.26	0.15	−0.40	−0.44	−0.46	−0.49
Incorrect Delta	−0.99	−0.99	0.15	−1.19	−1.24	−1.26	−1.31
(f) $t = 3$							

7. Methodological Assumptions and Implications in Practice

To apply the proposed two-phase RL approach to a hedging problem of contingent claims, there are at least four assumptions to be satisfied. This section discusses these assumptions and elaborates their implications in practice.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

7.1. Observable, sufficient, relevant, and transformed features in state

One of the crucial components in an MDP environment of the training phase or the online learning phase is the state, in which the features provide information from the environment to the RL agent. First, the features must be observable by the RL agent for learning. For instance, in our proposed state vectors (18) and (19), all the four features, namely the segregated account value, the hedging portfolio value, the number of surviving policyholders, and the term to maturity, are observable. Any unobservable, albeit desirable, features cannot be included in the state, such as insider information which could provide a better inference on the future value of a risky asset, or exact health condition of a policyholder. Second, the observable features in the state should be sufficient for the RL agent to learn. For example, due to the dual-risk bearing nature of the contract in this paper, the proposed state vectors (18) and (19) incorporate both financial and actuarial features; also, the third and the fourth features in the state vectors (18) and (19) would inform the RL agent to halt its hedging at the terminal time. However, incorporating sufficient observable features in the state does not imply that every observable feature in the environment should be included; the observable features in the state need to be relevant for learning efficiently. Since the segregated account value and the term to maturity have already been included in the state vectors (18) and (19) as features, the risky asset value and the hedging time are respective similar information from the environment and thus are redundant features to be contained in the state. Finally, the features in the state which have high variance might be appropriately transformed for reducing the volatility due to exploration. For instance, the segregated account value in the state vectors (18) and (19) is log-transformed in both phases.

7.2. Reward engineering

Another crucial component in an MDP environment is the reward, which supplies signals to the RL agent to evaluate its actions, i.e. the hedging strategy, for learning. First, the reward signals, if available, should suggest the local hedging performance. For example, in this paper, the RL agent is provided by the sequential anchor-hedging reward, given in (9), in the training phase; through the net liability value in the MDP training environment, the RL agent often receives a positive (resp. negative) signal for encouragement (resp. punishment), which is more informative than collecting the zero reward. However, any informative reward signals need to be computable from an MDP environment. In this paper, since the insurer does not know the MDP market environment, the RL agent could not be supplied the sequential anchor-hedging reward signals, which consist of the net liability values, in the online learning phase, even though they are more informative; instead, the RL agent is given the less informative single terminal reward, given in (8), in the online learning phase which can be computed from the market environment.

7.3. Markov property in state and action

In an MDP environment of the training phase or the online learning phase, the state and action pair needs to satisfy the Markov property as in (3). In the training phase, since the MDP training environment is constructed, the Markov property can be verified theoretically for the state, with the included features in line with section 7.1, and the action, which is the hedging strategy. For example, in this paper, with the model of the market environment being the BS and the CFM, the state vector in (18) and the Markovian hedging strategy satisfy the Markov property in the training phase. Since the illustrative example in this paper assumes that the market environment also follows the BS and the CFM, the state vector in (19) and the Markovian hedging strategy satisfy the Markov property in the online learning phase as well. However, in general, as the market environment is unknown, the Markov property for the state and action pair would need to be checked statistically in the online phase as follows.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

After the training phase and before an RL agent proceeding to the online learning phase, historical state and action sequences in a time frame are derived by hypothetically writing identical contingent claims and using the historical realisations from the market environment. For instance, historical values of risky assets are publicly available, or an insurer retrieves historical survival status of its policyholders with similar demographic information and medical history as the policyholder being actually written. These historical samples of the state and action pair are then used to conduct hypothesis testing on whether the Markov property in (3) holds for the pair in the market environment, by, for example, the test statistics proposed in Chen & Hong (2012). If the Markov property holds statistically, the RL agent could begin the online learning phase. Yet, if the property does not hold statistically, the state and action pair should be revised and then the training phase should be revisited; since the hedging strategy is the action in a hedging problem, only the state could be amended by including more features from the environment. Moreover, during the online learning phase, right after each further update step, new historical state and action sequences in a shifted time frame of the same duration are obtained together with the most recent historical realisations from the market environment and using the action samples being drawn from the updated policy. These regularly new samples should be applied to statistically verify the Markov property on a rolling basis. If the property fails to hold at any time, the state needs to be revised and the RL agent must be re-trained before resuming the online learning.

7.4. Re-establishment of contingent claims in online learning phase

Any contingent claims must have a finite terminal time realisation. On one hand, in the training phase, that would be the time when an episode ends and the state is re-initialised so that the RL agent can be trained in the training environment as long as possible. On the other hand, in the online learning phase, the market environment, and hence the state, could not be re-initialised; instead, at each terminal time realisation, the seller re-establishes identical contingent claims of the same contract characteristics and writing on (more or less) the same assets so that the RL agent can be trained in the market environment successively. In this paper, the terms to maturity and the minimum guarantees of all variable annuity contracts in the online learning phase are the same. Moreover, all re-established contracts therein write on the same financial risky asset, though the initial values of the asset are given by the real-time realisations in the market environment. Finally, while a new policyholder is written at each contract inception time, these policyholders have similar, if not identical, distributions of their random future lifetimes via examining their demographic information and medical history.

8. Concluding Remarks and Future Directions

This paper proposed the two-phase deep RL approach which can tackle practically common model miscalibration in hedging variable annuity contracts with both GMMB and GMDB riders in the BS financial and CFM actuarial market environments. The approach is composed of the training phase and the online learning phase. While the satisfactory hedging performance of the trained RL agent in the training environment was anticipated, the performance by the further trained RL agent in the market environment via the illustrative example should be highlighted. First, by comparing their sample means of terminal P&L from simulated scenarios, in most future trajectories, within a reasonable amount of time, the further trained RL agent was able to exceed the hedging performance by the correct Delta from the market environment and the incorrect Delta from the training environment. Second, through a more delicate hypothesis testing analysis, similar conclusions can be drawn in a fair amount of future trajectories. Finally, snapshots of empirical density functions, among the future trajectories, of the sample means of terminal P&L from simulated scenarios by each hedging strategy, shed light on the high probability that the further trained RL agent is indeed able to self-revise the hedging strategy.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

There should be at least two future directions derived from this paper. (I) The market environment in the illustrative example of this paper was assumed to be the BS financial and CFM actuarial models, which turned out to be the same as designed by the insurer for the training environment, with different parameters though. Moreover, the policyholders were assumed to be homogeneous that their survival probabilities and investment behaviours are all the same, with even identical contracts of the same minimum guarantee and maturity. In the market environment, the agent only had to hedge one contract at a time, instead of a portfolio of contracts. Obviously, if any of these is to be relaxed, the trained RL agent from the current training environment should not be able to produce satisfactory hedging performance in a market environment. Therefore, the training environment will certainly need to be substantially extended in terms of its sophistication, in order for the trained RL agent to be able to further learn and hedge well in any realistic market environments. (II) Beyond this, an even more ambitious question needs to be addressed is that how much similar do the training and market environments have to be, such that the online learning for self-revision on hedging strategy is possible, if not efficient. This second future direction is related to the transfer learning being adapted to the variable annuities hedging problem and shall be investigated carefully in the future.

References

Baydin, **A.G.**, **Pearlmutter**, **B.A.**, **Radul**, **A.A.** & **Siskind**, **J.M.** (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, **18**(2), 1–43.

Bertsimas, **D.**, **Kogan**, **L.** & **Lo**, **A.W.** (2000). When is time continuous? *Journal of Financial Economics*, **5**(2), 173–204.

Bühler, **H.**, **Gonon**, **L.**, **Teichmann**, **J.** & **Wood**, **B.** (2019). Deep hedging. *Quantitative Finance*, **19**(8), 1271–1291.

Cao, **J.**, **Chen**, **J.**, **Hull**, **J.** & **Poulos**, **Z.** (2021). Deep hedging of derivatives using reinforcement learning. *Journal of Financial Data Science*, **3**(1), 10–27.

Carbonneau, **A.** (2021). Deep hedging of long-term financial derivatives. *Insurance: Mathematics and Economics*, **99**, 327–340.

Charpentier, **A.**, **Élie**, **R.** & **Remlinger**, **C.** (2021). Reinforcement learning in economics and finance. *Computational Economics*.

Chen, **B.** & **Hong**, **Y.** (2012). Testing for the Markov property in time series. *Econometric Theory*, **28**, 130–178.

Chen, **Z.**, **Vetzal**, **K.** & **Forsyth**, **P.** (2008). The effect of modelling parameters on the value of GMWB guarantees. *Insurance: Mathematics and Economics*, **43**(1), 165–173.

Cheridito, **P.**, **Ery**, **J.** & **Wüthrich**, **M.V.** (2020). Assessing asset-liability risk with neural networks. *Risks*, **8**(1), article 16.

Chong, **W.F.** (2019). Pricing and hedging equity-linked life insurance contracts beyond the classical paradigm: the principle of equivalent forward preferences. *Insurance: Mathematics and Economics*, **88**, 93–107.

Cui, **Z.**, **Feng**, **R.** & **MacKay**, **A.** (2017). Variable annuities with VIX-linked fee structure under a Heston-type stochastic volatility model. *North American Actuarial Journal*, **21**(3), 458–483.

Dai, **M.**, **Kwok**, **Y.K.** & **Zong**, **J.** (2008). Guaranteed minimum withdrawal benefit in variable annuities. *Mathematical Finance*, **18**(4), 595–611.

Dang, **O.**, **Feng**, **M.** & **Hardy**, **M.R.** (2020). Efficient nested simulation for conditional tail expectation of variable annuities. *North American Actuarial Journal*, **24**(2), 187–210.

Dang, **O.**, **Feng**, **M.** & **Hardy**, **M.R.** (2022). Dynamic importance allocated nested simulation for variable annuity risk measurement. *Annals of Actuarial Science*, **16**(2), 319–348.

Feng, **B.M.**, **Tan**, **Z.** & **Zheng**, **J.** (2020). Efficient simulation designs for valuation of large variable annuity portfolios. *North American Actuarial Journal*, **24**(2), 275–289.

Feng, **R.** (2018). *An Introduction to Computational Risk Management of Equity-Linked Insurance*. CRC Press, Boca Raton, Florida, U.S.

Feng, **R.** & **Yi**, **B.** (2019). Quantitative modeling of risk management strategies: stochastic reserving and hedging of variable annuity guaranteed benefits. *Insurance: Mathematics and Economics*, **85**, 60–73.

Gan, **G.** (2013). Application of data clustering and machine learning in variable annuity valuation. *Insurance: Mathematics and Economics*, **53**(3), 795–801.

Gan, **G.** (2018). Valuation of large variable annuity portfolios using linear models with interactions. *Risks*, **6**(3), 1–19.

Gan, **G.** & **Lin**, **X.S.** (2015). Valuation of large variable annuity portfolios under nested simulation: a functional data approach. *Insurance: Mathematics and Economics*, **62**, 138–150.

Gan, **G.** & **Lin**, **X.S.** (2017). Efficient Greek calculation of variable annuity portfolios for dynamic hedging: a two-level metamodeling approach. *North American Actuarial Journal*, **21**(2), 161–177.

Gan, **G.** & **Valdez**, **E.A.** (2017). Modeling partial Greeks of variable annuities with dependence. *Insurance: Mathematics and Economics*, **76**, 118–134.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Gan, **G.** & **Valdez**, **E.A.** (2018). Regression modeling for the valuation of large variable annuity portfolios. *North American Actuarial Journal*, **22**(1), 40–54.

Gan, **G.** & **Valdez**, **E.A.** (2020). Valuation of large variable annuity portfolios with rank order kriging. *North American Actuarial Journal*, **24**(1), 100–117.

Gao, **G.** & **Wüthrich**, **M.V.** (2019). Convolutional neural network classification of telematics car driving data. *Risks*, **7**(1), article 6.

Gweon, **H.**, **Li**, **S.** & **Mamon**, **R.** (2020). An effective bias-corrected bagging method for the valuation of large variable annuity portfolios. *ASTIN Bulletin: The Journal of the International Actuarial Association*, **50**(3), 853–871.

Hardy, **M.** (2003). *Investment Guarantees: Modeling and Risk Management for Equity-Linked Life Insurance*. John Wiley & Sons, Inc., Hoboken, New Jersy, U.S.

Hasselt, **H.** (2010). Double Q-learning. *In Advances in Neural Information Processing Systems*, vol. **23**.

Hejazi, **S.A.** & **Jackson**, **K.R.** (2016). A neural network approach to efficient valuation of large portfolios of variable annuities. *Insurance: Mathematics and Economics*, **70**, 169–181.

Hu, **C.**, **Quan**, **Z.** & **Chong**, **W.F.** (2022). Imbalanced learning for insurance using modified loss functions in tree-based models. *Insurance: Mathematics and Economics*, **106**, 13–32.

Jeon, **J.** & **Kwak**, **M.** (2018). Optimal surrender strategies and valuations of path-dependent guarantees in variable annuities. *Insurance: Mathematics and Economics*, **83**, 93–109.

Kindratenko, **V.**, **Mu**, **D.**, **Zhan**, **Y.**, **Maloney**, **J.**, **Hashemi**, **S.H.**, **Rabe**, **B.**, **Xu**, **K.**, **Campbell**, **R.**, **Peng**, **J.** & **Gropp**, **W.** (2020). HAL: computer system for scalable deep learning. In *Practice and Experience in Advanced Research Computing* (PEARC'20) (pp. 41–48).

Kolm, **P.N.** & **Ritter**, **G.** (2019). Dynamic replication and hedging: a reinforcement learning approach. *Journal of Financial Data Science*, **1**(1), 159–171.

Lin, **X.S.** & **Yang**, **S.** (2020). Fast and efficient nested simulation for large variable annuity portfolios: a surrogate modeling approach. *Insurance: Mathematics and Economics*, **91**, 85–103.

Liu, **K.** & **Tan**, **K.S.** (2020). Real-time valuation of large variable annuity portfolios: a green mesh approach. *North American Actuarial Journal*, **25**(3), 313–333.

Milevsky, **M.A.** & **Posner**, **S.E.** (2001). The Titanic option: valuation of the guaranteed minimum death benefit in variable annuities and mutual funds. *The Journal of Risk and Insurance*, **68**(1), 93–128.

Milevsky, **M.A.** & **Salisbury**, **T.S.** (2006). Financial valuation of guaranteed minimum withdrawal benefits. *Insurance: Mathematics and Economics*, **38**(1), 21–38.

Mnih, **V.**, **Kavukcuoglu**, **K.**, **Silver**, **D.**, **Graves**, **A.**, **Antonoglou**, **I.**, **Wierstra**, **D.** & **Riedmiller**, **M.** (2013). Playing Atari with deep reinforcement learning. arXiv: 1312.5602.

Mnih, **V.**, **Kavukcuoglu**, **K.**, **Silver**, **D.**, **Rusu**, **A.A.**, **Veness**, **J.**, **Bellemare**, **M.G.**, **Graves**, **A.**, **Riedmiller**, **M.**, **Fidjeland**, **A.K.**, **Ostrovski**, **G.**, **Petersen**, **S.**, **Beattie**, **C.**, **Sadik**, **A.**, **Antonoglou**, **I.**, **King**, **H.**, **Kumaran**, **D.**, **Wierstra**, **D.**, **Legg**, **S.** & **Hassabis**, **D.** (2015). Human-level control through deep reinforcement learning. *Nature*, **518**, 529–533.

Moenig, **T.** (2021a). Efficient valuation of variable annuity portfolios with dynamic programming. *Journal of Risk and Insurance*, **88**(4), 1023–1055.

Moenig, **T.** (2021b). Variable annuities: market incompleteness and policyholder behavior. *Insurance: Mathematics and Economics*, **99**, 63–78.

Perla, **F.**, **Richman**, **R.**, **Scognamiglio**, **S.** & **Wüthrich**, **M.V.** (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, **7**, 572–598.

Quan, **Z.**, **Gan**, **G.** & **Valdez**, **E.** (2021). Tree-based models for variable annuity valuation: parameter tuning and empirical analysis. *Annals of Actuarial Science*, **16**(1), 95–118.

Richman, **R.** & **Wüthrich**, **M.V.** (2021). A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science*, **15**(2), 346–366.

Schulman, **J.**, **Levine**, **S.**, **Moritz**, **P.**, **Jordan**, **M.** & **Abbeel**, **P.** (2015). Trust region policy optimization. arXiv: 1502.05477.

Schulman, **J.**, **Wolski**, **F.**, **Dhariwal**, **P.**, **Radford**, **A.** & **Klimov**, **O.** (2017). Proximal policy optimization algorithms. arXiv: 1707.06347.

Silver, **D.**, **Schrittwieser**, **J.**, **Simonyan**, **K.**, **Antonoglou**, **I.**, **Huang**, **A.**, **Guez**, **A.**, **Hubert**, **T.**, **Baker**, **L.**, **Lai**, **M.**, **Bolton**, **A.**, **Chen**, **Y.**, **Lillicrap**, **T.**, **Hui**, **F.**, **Sifre**, **L.**, **Van den Driessche**, **G.**, **Graepel**, **T.** & **Hassabis**, **D.** (2017). Mastering the game of Go without human knowledge. *Nature*, B, 354–359.

Sutton, **R.S.** (1984). *Temporal Credit Assignment in Reinforcement Learning*. PhD thesis, University of Massachusetts.

Sutton, **R.S.** (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, **3**, 9–44.

Sutton, **R.S.** & **Barto**, **A.G.** (2018). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts, U.S.

Trottier, **D.A.**, **Godin**, **F.** & **Hamel**, **E.** (2018). Local hedging of variable annuities in the presence of basis risk. *ASTIN Bulletin: The Journal of the International Actuarial Association*, **48**(2), 611–646.

Wang, **G.** & **Zou**, **B.** (2021). Optimal fee structure of variable annuities. *Insurance: Mathematics and Economics*, **101**, 587–601.

Wang, **H.**, **Zariphopoulou**, **T.** & **Zhou**, **X.** (2020). Reinforcement learning in continuous time and space: a stochastic control approach. *Journal of Machine Learning Research*, **21**, 1–34.

Wang, **H.** & **Zhou**, **X.** (2020). Continuous-time mean-variance portfolio selection: a reinforcement learning framework. *Mathematical Finance*, **30**(4), 1273–1308.
Watkins, **C.J.C.H.** (1989). *Learning from Delayed Rewards*. PhD thesis, University of Cambridge.
Watkins, **C.J.C.H.** & **Dayan**, **P.** (1992). Q-learning. *Machine Learning*, **8**, 297–292.
Weaver, **L.** & **Tao**, **N.** (2001). The Optimal Reward Baseline for Gradient-Based Reinforcement Learning. UAI'01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, 538–545.
Williams, **R.J.** (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, **8**, 229–256.
Wüthrich, **M.V.** (2018). Neural networks applied to chain-ladder reserving. *European Actuarial Journal*, **8**, 407–436.
Xu, **W.**, **Chen**, **Y.**, **Coleman**, **C.** & **Coleman**, **T.F.** (2018). Moment matching machine learning methods for risk management of large variable annuity portfolios. *Journal of Economic Dynamics and Control*, **87**, 1–20.
Xu, **X.** (2020). *Variable Annuity Guaranteed Benefits: An Integrated Study of Financial Modelling, Actuarial Valuation and Deep Learning*. PhD thesis, UNSW Business School.

Appendix A. Deep Hedging Approach

In this section, we provide a brief review of the DH approach adapted from Bühler *et al.* (2019). In particular, the hedging objective of the insurer is still given as $\sqrt{\mathbb{E}\left[\left(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\right)^2\right]}$, with Equation (2) being the optimal (discrete) hedging strategy. The hedging agent built by the insurer using the DH algorithm shall be called the DH agent hereafter.

A.1. Deterministic Action

Different from section 3.1, in which the RL agent takes a stochastic action which is sampled from the policy for the exploration in the MDP environment, the DH agent only deploys a deterministic action $H^{\mathrm{DH}} : \mathcal{X} \to \mathcal{A}$, which is a direct mapping from the state space to the action space. Specifically, at each time t_k, where $k = 0, 1, \ldots, n-1$, given the current state $X_{t_k} \in \mathcal{X}$, the DH agent takes an action $H^{\mathrm{DH}}\left(X_{t_k}\right) \in \mathcal{A}$. In this case, the objective of the DH agent is to solve for the optimal hedging strategy $H^{\mathrm{DH},*}(\cdot)$ that minimises $\sqrt{\mathbb{E}\left[\left(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\right)^2\right]}$, or equivalently minimises $\mathbb{E}\left[\left(P_{t_{\tilde{n}}} - L_{t_{\tilde{n}}}\right)^2\right]$.

A.2. Action Approximation and Parameterisation

The deterministic action mapping $H^{\mathrm{DH}} : \mathcal{X} \to \mathcal{A}$ is then approximated and parameterised by an ANN with weights υ_{a}. The construction of such ANN $\mathcal{N}_{\mathrm{a}}(\cdot; \upsilon_{\mathrm{a}})$ is similar to that in section 3.3.1, except that $\mathcal{N}_{\mathrm{a}}(x; \upsilon_{\mathrm{a}}) \in \mathbb{R}$ for any $x \in \mathbb{R}^p$; that is, $\mathcal{N}_{\mathrm{a}}(\cdot; \upsilon_{\mathrm{a}})$ takes a state vector $x \in \mathbb{R}^p$ as the input, and directly outputs a deterministic action $a\,(x; \upsilon_{\mathrm{a}}) \in \mathbb{R}$, instead of the Gaussian mean-variance tuple $\left(c(x; \upsilon_{\mathrm{a}})\,, d^2\,(x; \upsilon_{\mathrm{a}})\right) \in \mathbb{R} \times \mathbb{R}^+$ in the RL approach, which then samples an action from the Gaussian measure. Hence, in the DH approach, solving the optimal hedging strategy $H^{\mathrm{DH},*}(\cdot)$ boils down to finding the optimal weights υ_a^*.

A.3. Deep Hedging Method

The DH agent starts from initial ANN weights $\upsilon_{\mathrm{a}}^{(0)}$, deploys the hedging strategy to collect terminal P&Ls, and gradually updates the ANN weights by stochastic gradient ascent as shown in Equation (13), with θ replaced by υ. For the DH agent, at each update step $u = 1, 2, \ldots$, the surrogate performance measure is given as

$$\mathcal{J}^{(u-1)}\left(\upsilon_{\mathrm{a}}^{(u-1)}\right) = -\mathbb{E}\left[\left(P_{t_{\tilde{n}}}^{(u-1)} - L_{t_{\tilde{n}}}^{(u-1)}\right)^2\right].$$

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

Correspondingly, the gradient of the surrogate performance measure with respect to the ANN weights υ_{a} is

$$\nabla_{\upsilon_{\mathrm{a}}}\mathcal{J}^{(u-1)}\left(\upsilon_{\mathrm{a}}^{(u-1)}\right)=-2\mathbb{E}\left[\left(P_{t_{\tilde{n}}}^{(u-1)}-L_{t_{\tilde{n}}}^{(u-1)}\right)\nabla_{\upsilon_{\mathrm{a}}}P_{t_{\tilde{n}}}^{(u-1)}\right].$$

Therefore, based on the *realised* terminal P&L $p_{t_{\tilde{n}}}^{(u-1)}$ and $l_{t_{\tilde{n}}}^{(u-1)}$, the estimated gradient is given as

$$\widehat{\nabla_{\upsilon_{\mathrm{a}}}\mathcal{J}^{(u-1)}\left(\upsilon_{\mathrm{a}}^{(u-1)}\right)}=-2\left(p_{t_{\tilde{n}}}^{(u-1)}-l_{t_{\tilde{n}}}^{(u-1)}\right)\nabla_{\upsilon_{\mathrm{a}}}p_{t_{\tilde{n}}}^{(u-1)}.$$

Algorithm 1 summarises the DH method above.

Algorithm 1. Pseudo-code for deep hedging method

Input initial ANN model $\mathcal{N}_{\mathrm{a}}\left(\cdot;\upsilon_{\mathrm{a}}^{(0)}\right)$, total number of updates $\hat{M}\in\mathbb{N}$, learning rate $\hat{\alpha}\in[0,1]$.

for $u=1,2,\cdots,\hat{M}$ **do**

- Initialise the MDP training environment and observe the initial state vector $x_{t_0}^{(u-1)}$.
- Follow the hedging strategy $\mathcal{N}_{\mathrm{a}}\left(\cdot;\upsilon_{\mathrm{a}}^{(u-1)}\right)$ to realise an episode and evaluate the terminal P&L $p_{t_{\tilde{n}}}^{(u-1)}$ and $l_{t_{\tilde{n}}}^{(u-1)}$.
- Update $\upsilon_{\mathrm{a}}^{(u-1)}$ as

$$\upsilon_{\mathrm{a}}^{(u)}=\upsilon_{\mathrm{a}}^{(u-1)}-2\hat{\alpha}\left(p_{t_{\tilde{n}}}^{(u-1)}-l_{t_{\tilde{n}}}^{(u-1)}\right)\nabla_{\upsilon_{\mathrm{a}}}p_{t_{\tilde{n}}}^{(u-1)}.$$

end

Return the trained ANN model $\mathcal{N}_{\mathrm{a}}\left(\cdot;\upsilon_{\mathrm{a}}^{(\hat{M})}\right)$.

Compared with policy gradient methods introduced in section 3.4, the DH method shows two key differences. First, it assumes that the hedging portfolio value $P_{t_{\tilde{n}}}^{(u-1)}$ is differentiable with respect to υ_{a} at each update $u=1,2,\ldots$. Second, the update of ANN weights does not depend on intermediate rewards collected during an episode; that is, to update the weights, the DH agent has to experience a complete episode to realise the terminal P&L. Therefore, the update frequency of the DH method is lower than that of the RL method with TD feature.

Appendix B. REINFORCE: A Monte Carlo Policy Gradient Method

At each update step $u=1,2,\ldots,$ based on the ANN weights $\theta^{(u-1)}$, and thus the policy $\pi\left(\cdot;\theta_{\mathrm{p}}^{(u-1)}\right)$, the RL agent experiences the realised episode:

$$\left\{x_{t_0}^{(u-1)},h_{t_0}^{(u-1)},x_{t_1}^{(u-1)},r_{t_1}^{(u-1)},h_{t_1}^{(u-1)},\ldots,x_{t_{\tilde{n}-1}}^{(u-1)},r_{t_{\tilde{n}-1}}^{(u-1)},h_{t_{\tilde{n}-1}}^{(u-1)},x_{t_{\tilde{n}}}^{(u-1)},r_{t_{\tilde{n}}}^{(u-1)}\right\},$$

where $h_{t_k}^{(u-1)}$, for $k=0,1,\ldots,\tilde{n}-1$, is the time t_k realised hedging strategy being sampled from the Gaussian distribution with the mean $c\left(x_{t_k}^{(u-1)};\theta_{\mathrm{p}}^{(u-1)}\right)$ and the variance $d^2\left(x_{t_k}^{(u-1)};\theta_{\mathrm{p}}^{(u-1)}\right)$. In the following, fix an update step $u=1,2,\ldots$.

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

REINFORCE takes directly the time-0 value function $V^{(u-1)}\big(0,x;\theta_{\mathrm{p}}\big)$, for any $x\in\mathcal{X}$, as a part of the surrogate performance measure:

$$V^{(u-1)}\big(0,x;\theta_{\mathrm{p}}\big)=\mathbb{E}\left[\sum_{k=0}^{\tilde{n}-1}R_{t_{k+1}}^{(u-1)}\Big|X_0^{(u-1)}=x\right].$$

In Williams (1992), the *Policy Gradient Theorem* was proved, which states that

$$\nabla_{\theta_{\mathrm{p}}}V^{(u-1)}\big(0,x;\theta_{\mathrm{p}}\big)=\mathbb{E}\left[\sum_{k=0}^{\tilde{n}-1}\left(\sum_{l=k}^{\tilde{n}-1}R_{t_{l+1}}^{(u-1)}\right)\nabla_{\theta_{\mathrm{p}}}\ln\phi\left(H_{t_k}^{(u-1)};X_{t_k}^{(u-1)},\theta_{\mathrm{p}}\right)\Big|X_0^{(u-1)}=x\right],$$

where $\phi\left(\cdot;X_{t_k}^{(u-1)},\theta_{\mathrm{p}}\right)$ is the Gaussian density function with mean $c\left(X_{t_k}^{(u-1)};\theta_{\mathrm{p}}\right)$ and variance $d^2\left(X_{t_k}^{(u-1)};\theta_{\mathrm{p}}\right)$. Therefore, based on the realised episode, the estimated gradient of the time-0 value function is given by

$$\widehat{\nabla_{\theta_{\mathrm{p}}}V^{(u-1)}\left(0,x;\theta_{\mathrm{p}}^{(u-1)}\right)}=\sum_{k=0}^{\tilde{n}-1}\left(\sum_{l=k}^{\tilde{n}-1}r_{t_{l+1}}^{(u-1)}\right)\nabla_{\theta_{\mathrm{p}}}\ln\phi\left(h_{t_k}^{(u-1)};x_{t_k}^{(u-1)},\theta_{\mathrm{p}}^{(u-1)}\right).$$

Notice that, thanks to the Policy Gradient Theorem, the gradient of the surrogate performance measure does not depend on the gradient of the reward function, and hence, the reward function could be discrete or non-differentiable while the estimated gradient of the surrogate performance measure only needs the numerical reward values. However, in the DH approach of Bühler *et al.* (2019), the gradient of the surrogate performance measure therein does depend on the gradient of the terminal loss function and thus that approach implicitly requires the differentiability of the hedging portfolio value while the estimated gradient of the surrogate performance requires its numerical gradient values. See Appendix A for more details.

To reduce the variance of estimated gradient above, Williams (1992) suggested to introduce an unbiased baseline in this gradient, where a natural choice is the value function:

$$\nabla_{\theta_{\mathrm{p}}}V^{(u-1)}\big(0,x;\theta_{\mathrm{p}}\big)=\mathbb{E}\left[\sum_{k=0}^{\tilde{n}-1}\left(\sum_{l=k}^{\tilde{n}-1}R_{t_{l+1}}^{(u-1)}-V\left(t_k,X_{t_k}^{(u-1)};\theta_{\mathrm{p}}\right)\right)\right.$$
$$\left.\nabla_{\theta_{\mathrm{p}}}\ln\phi\left(H_{t_k}^{(u-1)};X_{t_k}^{(u-1)},\theta_{\mathrm{p}}\right)\Big|X_0^{(u-1)}=x\right];$$

see also Weaver and Tao (2001). Herein, at any time t_k, for $k=0,1,\ldots,\tilde{n}-1$, $A_{t_k}^{(u-1)}=\sum_{l=k}^{\tilde{n}-1}R_{t_{l+1}}^{(u-1)}-V\left(t_k,X_{t_k}^{(u-1)};\theta_{\mathrm{p}}\right)$ is called an *advantage*. Since the true value function is unknown to the RL agent, it is approximated by $\hat{V}\left(t_k,X_{t_k}^{(u-1)};\theta_{\mathrm{v}}^{(u-1)}\right)=\mathcal{N}_{\mathrm{v}}\left(X_{t_k}^{(u-1)};\theta_{\mathrm{v}}^{(u-1)}\right)$, defined in (12), and in which the ANN weights are evaluated at $\theta_{\mathrm{v}}=\theta_{\mathrm{v}}^{(u-1)}$ as the gradient of the time-0 value function is independent of the ANN weights θ_{v}; hence, the estimated advantage is given by $\hat{A}_{t_k}^{(u-1)}=\sum_{l=k}^{\tilde{n}-1}R_{t_{l+1}}^{(u-1)}-\hat{V}\left(t_k,X_{t_k}^{(u-1)};\theta_{\mathrm{v}}^{(u-1)}\right)$.

Due to the value function approximation in the baseline, REINFORCE includes a second component in the surrogate performance measure, which aims to minimise the loss between the sum of reward signals and the approximated value function by the ANN. Therefore, the surrogate performance measure is given by:

https://doi.org/10.1017/S1748499523000027 Published online by Cambridge University Press

$$\mathcal{J}^{(u-1)}(\theta) = V^{(u-1)}\left(0, x; \theta_{\mathrm{p}}\right) - \mathbb{E}\left[\sum_{k=0}^{\tilde{n}-1}\left(\hat{A}^{(u-1)}_{\theta_{\mathrm{p}}^{(u-1)}, t_k} + \hat{V}\left(t_k, X^{(u-1)}_{t_k}; \theta_{\mathrm{v}}^{(u-1)}\right)\right.\right.$$
$$\left.\left. - \hat{V}\left(t_k, X^{(u-1)}_{t_k}; \theta_{\mathrm{v}}\right)\right)^2 \middle| X^{(u-1)}_0 = x\right],$$

where the estimated advantaged $\hat{A}^{(u-1)}_{\theta_{\mathrm{p}}^{(u-1)}, t_k}$ is evaluated at $\theta_{\mathrm{p}} = \theta_{\mathrm{p}}^{(u-1)}$.

Hence, at each update step $u = 1, 2, \ldots$, based on the ANN weights $\theta^{(u-1)}$, and thus, the policy $\pi\left(\cdot; \theta_{\mathrm{p}}^{(u-1)}\right)$, the estimated gradient of the surrogate performance measure is given by

$$\begin{aligned}
\nabla_\theta \widehat{\mathcal{J}^{(u-1)}\left(\theta^{(u-1)}\right)} &= \sum_{k=0}^{\tilde{n}-1}\left(\sum_{l=k}^{\tilde{n}-1} r^{(u-1)}_{t_{l+1}} - \hat{V}\left(t_k, x^{(u-1)}_{t_k}; \theta_{\mathrm{v}}^{(u-1)}\right)\right) \nabla_{\theta_{\mathrm{p}}} \ln \phi\left(h^{(u-1)}_{t_k}; x^{(u-1)}_{t_k}, \theta_{\mathrm{p}}^{(u-1)}\right) \\
&\quad + \sum_{k=0}^{\tilde{n}-1}\left(\sum_{l=k}^{\tilde{n}-1} r^{(u-1)}_{t_{l+1}} - \hat{V}\left(t_k, x^{(u-1)}_{t_k}; \theta_{\mathrm{v}}^{(u-1)}\right)\right) \nabla_{\theta_{\mathrm{v}}} \hat{V}\left(t_k, x^{(u-1)}_{t_k}; \theta_{\mathrm{v}}^{(u-1)}\right) \\
&= \sum_{k=0}^{\tilde{n}-1} \hat{a}^{(u-1)}_{t_k}\left(\nabla_{\theta_{\mathrm{p}}} \ln \phi\left(h^{(u-1)}_{t_k}; x^{(u-1)}_{t_k}, \theta_{\mathrm{p}}^{(u-1)}\right) + \nabla_{\theta_{\mathrm{v}}} \hat{V}\left(t_k, x^{(u-1)}_{t_k}; \theta_{\mathrm{v}}^{(u-1)}\right)\right),
\end{aligned}$$

where $\hat{a}^{(u-1)}_{t_k} = \sum_{l=k}^{\tilde{n}-1} r^{(u-1)}_{t_{l+1}} - \hat{V}\left(t_k, x^{(u-1)}_{t_k}; \theta_{\mathrm{v}}^{(u-1)}\right)$, for $k = 0, 1, \ldots, \tilde{n}-1$, is the realised estimated advantage.

Appendix C. Deep Hedging Training

The state vector observed by the DH agent is the same as that by the RL agent in Equation (18). Table C.1(a) summarises the hyperparameters of DH agent training, while Table C.1(b) outlines the hyperparameters of the ANN architecture of DH agent; see Appendix A.

Table C.1. The hyperparameters of deep hedging training and the neural network.

(a) Hyperparameters of deep hedging training	
Parameter	Value
Number of updates $\hat{M}$	10^8
Learning rate $\hat{\alpha}$	0.0001
Optimiser	Adam
(b) Hyperparameters for neural network	
Parameter	Value(s)
Number of layers	6
Dimension of hidden layers	[32, 64, 128, 64, 32]
Activation function	ReLU

Cite this article: Chong WF, Cui H and Li Y (2023). Pseudo-model-free hedging for variable annuities via deep reinforcement learning, *Annals of Actuarial Science*, **17**, 503–546. https://doi.org/10.1017/S1748499523000027