

# Optimal strategies for options on target volatility funds\*

Roberto Daluiso<sup>1</sup>, Emanuele Nastasi<sup>2</sup>, Andrea Pallavicini<sup>1</sup>, and Stefano Polo<sup>3</sup>

<sup>1</sup> Financial Engineering, Intesa Sanpaolo Group,  
Largo Mattioli 3, 20121 Milan, Italy

{Roberto.Daluiso, Andrea.Pallavicini}@intesasanpaolo.com

<sup>2</sup> Capital Market Technologies, swissQuant Group AG,  
Kuttelgasse 7, 8001 Zürich, Switzerland  
Nastasi@swissquant.com

<sup>3</sup> Financial Engineering, X-Numeris S.r.l.,  
Via Turro, 7, 20127 Milan, Italy  
Sp.Stefano.Polo@gmail.com

**Abstract.** We deal with a stochastic optimal control problem rising from hedging the risky securities underlying a target volatility strategy, a portfolio whose asset-allocation is adjusted to maintain the realized volatility of the portfolio at a certain level. We consider the point of view of a derivative writer selling an option contract as protection to a portfolio manager on the invested capital. The uncertainty in the risky portfolio composition along with the difference in hedging costs of its components requires to adjust the protection price to include these costs in the worst-case scenario for the seller. We derive an analytical solution of the problem in a Black and Scholes scenario. Then, we use Reinforcement Learning techniques to determine the fund composition leading to the optimal policy under the local volatility model, for which an *a priori* solution is not available. We show how the performances of the numerical solution are compatible with those obtained by applying path-wise the analytical solution previously derived.

**Keywords.** Stochastic control problem, reinforcement learning, target volatility, option pricing, hedging costs.

**M.S.C. classification.** 65C05, 68T07, 91G20.

**J.E.L. classification.** C63, C45, G13.

---

\* Our sincere thanks go to Marco Bianchetti and Diego Pierluigi Giovannini of Intesa Sanpaolo Milan, as well as to the Italian computing center Cineca for its support on our numerical simulations. The opinions here expressed are solely those of the authors and do not represent in any way those of their employers.

## 1 Introduction

In the recent years portfolio managers were exposed to very low interest rates and quickly changing market volatilities. An effective solution to control risks under such an environment is given by target volatility strategies (TVSs), also known as constant volatility targeting, which are able to preserve the portfolio at a predetermined level of volatility. A TVS is a strategy for managing a portfolio of risky assets (typically equities) and a risk-free one dynamically re-balanced with the aim of maintaining the overall portfolio volatility level close to some target value. The strategy is designed to achieve a stable level of volatility for the underlying portfolio in any market scenario, and to allow a free selection of the relative allocation weights among the risky assets. The constant volatility approach can help investors to obtain desired risk exposures over the short and long term and possibly it increases the risk-adjusted performance of the portfolio.

Financial products embedding this strategy were initially offered in the Asian markets, see for instance [3] and [28], which highlight the pros and cons for investors, to be adopted in the following years in many other markets in North America and Europe as depicted in [21]. At the present day, we can observe some new market indices based on the mechanism of the target volatility strategy such as Dow Jones Volatility Control Index, and S&P 500 Risk Control Index. In the recent literature, TVS-based portfolios are investigated with respect to their performances in terms of realized returns, see for instance [17] and [22], and the soundness of the volatility targeting algorithm, as described in [18], where the authors propose the use of artificial neural networks for volatility forecasting to enhance the performance of an asset allocation strategy.

The success of TVS-based products has led to the emergence of derivative contracts, known as target volatility options (TVO), where the TVS itself is used as underlying asset of the contract. In the derivative pricing literature TVOs are reviewed in [1], [6], [7], and [13].

This paper aims to enrich the pricing framework of the TVOs by studying for the first time, to the best of our knowledge, the funding costs coming from hedging (or simply hedging costs) the risky assets underlying the target volatility strategy. In particular, we consider the point of view of an option writer (for example a bank) selling a call option to a portfolio manager as protection on the capital invested in a TVS. During her activity the portfolio manager has the freedom of changing the relative weights of the risky assets during the life of the TVS depending on how the market evolves. Since the fund manager's decisions cannot be known *a priori* and since the risky assets have different hedging costs, the option writer shall adjust the price of the protection to include these costs in the worst-case scenario, i.e. the most expensive strategy in term of the hedging costs. Hence, the pricing problem becomes a continuous dynamical control problem over the risky portfolio composition: the problem of finding the manager's optimal strategy that leads to the maximum protection price.

In our work, we provide the reader with a formal description of the stochastic control problem and we show that it can be solved only numerically considering a general state-dependent dynamics for the risky assets. However, we prove that an

analytical solution to the problem exists, assuming that the risky asset, dynamics underlying the TVS portfolio follow a Black and Scholes (BS) model [2] and that the derivative contract is a European-style option. As our first contribution, we derive the BS closed solution in two different ways: either applying the Gyöngy Lemma or the Hamilton-Jacobi-Bellman equation.

In the second part of our work we tackle the problem in the general case of a state-dependent dynamics for the risky assets for which only the numerical approach is admissible. More precisely we numerically study the problem assuming a local volatility (LV) dynamics [5], [9] for the risky assets. The reason for selecting the LV model is its widespread adoption in the industry to value derivatives, especially those related to asset baskets [11]. The LV investigation is based on a Reinforcement Learning approach due to the high dimensionality of the problem we deal with. In particular, we compare the performances of a plain direct policy search gradient-based algorithm [27] with those of the more sophisticated proximal policy optimization technique developed in [25]. At the end of our study, we show that the solution found in the BS scenario provides a good approximation of the LV solution, which could be favored by practitioners if computational time is valued.

The paper is organized as follows. In Section 2 we introduce more in detail the contract between the option-writer and the manager and provide the description of the TVS dynamics in presence of valuation adjustments such as the hedging costs. Then, in Section 3 we introduce the structured class of derivative contracts linked to TVSs, where we describe the arising dynamical control problem for pricing those options. Moreover, in this section we derive the BS closed solution for European TVOs in two different ways: one applying the Gyöngy Lemma and the other through the Hamilton-Jacobi-Bellman equation. In Section 4 we illustrate how we have applied RL to solve the dynamic control problem, giving a description of the algorithm we have built. We conclude the paper with Section 5 where we present the numerical results obtained in this work for the Black and Scholes model and the local volatility one.

## 2 Target volatility strategy

As mentioned before, this work aims to enrich the TVS pricing literature by studying the aspect related to the funding costs coming from hedging the underlying risky assets. In the following, we will refer to these costs simply as hedging costs. We consider the following scenario: an option writer selling a protection to a portfolio manager who has her capital invested according to a TVS. In our case the fund manager has the freedom of changing the relative weights of the risky asset during the life of the TVS; once the allocation strategy is selected then the volatility targeting algorithm rebalances the risky component of the portfolio with the risk-free one in order to keep the overall portfolio volatility close to a target value. Clients investing in the fund pay a running fee for the service of the fund manager and their capital is protected.

The fund manager usually buys from a counterpart an option on the TVS to ensure capital protection. For instance, the capital can be protected by buying a put option. In this case, we can write the undiscounted net asset value (NAV)  $A_t$  of the strategy as given by<sup>1</sup>

$$A_t := \max\{V_t, K\} = V_t + (K - V_t)^+, \quad (1)$$

where  $V_t$  is the price process of the strategy, and  $K$  is the guaranteed capital. On the other hand, the fund manager can replicate the payoff by means of the put-call parity by investing the capital in a low-risk asset and buying a call on the strategy

$$A_t = K + (V_t - K)^+. \quad (2)$$

In this way, the TVS is only defined in the two contracts client-fund and fund-option writer. The fund manager is not implementing the strategy by trading in the market, and she is not subject to additional costs to access the market. On the other way, the writer is paying such costs since she is actively hedging the call option sold to the manager. The writer trading activity implemented to hedge the option requires funding the collateral procedures of the hedging instruments along with any lending/borrowing fee. We refer to [10] for a discussion of how to explicitly calculate hedging costs in equity trading.

We remark that, upon entering into the contract, the choices of the manager trading activity can be seen as stochastic processes since they are not known *a priori* and will depend on the market evolution. Thus, neither the manager strategy nor the writer hedging costs can be written in the fund-writer contract: the only contractual elements are the protection strike and maturity, the market in which the manager can trade and the target volatility of the TVS portfolio. For this reason, the price of a financial product sold by the writer is adjusted to include any valuation adjustment due to the trading activity, such as the hedging costs.

Our aim is to find the most expensive investing strategy from the point of view of hedging costs for the writer that the manager could choose in the market. In other words, we want to determine the worst-case scenario for the option-seller. In the current section, we proceed by defining the price process of the TVS so that we can highlight the impact of valuation adjustments.

## 2.1 The strategy price process

We work on a filtered probability space  $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbb{P})$  satisfying the usual assumptions for a market model, where  $\mathbb{P}$  is the physical probability measure.

We consider a fund trading a basket of  $n$  risky securities with price process  $S_t = \{S_t^i, i = 1, \dots, n\}$  funded with a cash account  $B_t$  accruing at  $r_t$ . Any dividend paid by the securities is re-invested in the fund. Here, we assume that the TVS is implemented in continuous time, even if in the practice we can

---

<sup>1</sup> Here we neglect discounting factors for sake of clarity.

implement the strategy only on a discrete set of dates. We introduce the deflated gain process  $\bar{G}_t^i$  associated with the risky securities as given by

$$\bar{G}_t^i := \bar{S}_t^i + \bar{D}_t^i, \quad (3)$$

where we define the deflated price and cumulative dividend processes as

$$\bar{S}_t^i := \frac{S_t^i}{B_t}, \quad \bar{D}_t^i := \int_0^t \frac{d\pi_u^i}{B_u} + \int_0^t \frac{d\psi_u^i}{B_u}, \quad (4)$$

where  $\pi_t^i$  represents the cumulative contractual-coupon process paid by the security, and  $\psi_t^i$  represents the cumulative valuation adjustments. Since fund managers allocating TVS usually rely on equity assets, here we use the results of [10] that analyze the valuation adjustments for equity products. We can write

$$\psi_t^i := \int_0^t S_u^i \mu_u^i du, \quad (5)$$

where we call  $\mu_t^i$  cost of carry, which basically represents the hedging costs for the  $i$ -th security.

Then, we introduce the strategy price process  $V_t$ , and we define the deflated gain process  $\bar{G}_t^V$  as

$$\bar{G}_t^V := \frac{V_t}{B_t} + \int_0^t \frac{V_u \phi_u}{B_u} du, \quad (6)$$

where  $\phi_t$  are the running fees earned by the fund manager for her activity. We assume that the strategy is self-financing, so that we can write

$$d\bar{G}_t^V = q_t \cdot d\bar{G}_t, \quad (7)$$

where  $q_t^i$  is the quantity invested in the  $i$ -th security.<sup>2</sup>

Now, in order to prevent arbitrages, we assume the existence of a risk-neutral measure  $\mathbb{Q}$  equivalent to  $\mathbb{P}$  under which the deflated gain processes of all traded securities are martingales. Under this assumption, we are able to derive the drift conditions on the security price processes, and in turn on the strategy price process, that is

$$\forall u > t \quad \bar{G}_t^i = \mathbb{E}_t [\bar{G}_u^i] \implies dS_t^i = r_t S_t^i dt - d\pi_t^i - d\psi_t^i + dM_t^i, \quad (8)$$

where  $M_t^i$  are martingales under  $\mathbb{Q}$ . If we substitute this expression for the security dynamics into the definition of the strategy we can check that the price process of the strategy is accruing at the following cash account rate  $r_t$  compensated for the fund manager fees

$$dV_t = V_t(r_t - \phi_t)dt + dM_t^V, \quad (9)$$

<sup>2</sup> In all formulae we use dot notation for scalar product between vectors, i.e.  $a \cdot b = \sum_i a_i b_i$ , or between matrix and vector, i.e.  $A \cdot b = \sum_j a_{ij} b_j$  or  $b \cdot A = \sum_i b_i a_{ij}$ .

with  $M_t^V$  martingale under  $\mathbb{Q}$ . Notice that, as expected from non-arbitrage considerations, the coupons paid by each security appear only in the drift of the security price process, but they do not impact the drift of the strategy.

Yet, the strategy priced by  $V_t$  cannot be described in the contract between the parties, since Equation (7) depends via the security gain processes on the valuation adjustment  $\psi_t^i$ , which is specific to the investor pricing the strategy. Thus, the TVS defined in the contract will be

$$d\bar{I}_t := q_t \cdot \left( d\bar{S}_t + \frac{d\pi_t}{B_t} \right) - \bar{I}_t \phi_t dt \quad \text{with} \quad I_0 = V_0, \quad (10)$$

leading to the following price process dynamics

$$dI_t = I_t(r_t - \phi_t)dt - q_t \cdot d\psi_t + dM_t^I, \quad (11)$$

with  $M_t^I$  martingale under  $\mathbb{Q}$ . In this case, we observe that  $I_t$  depends explicitly both on the valuation adjustments and on the allocation strategy. Indeed, if we substitute the valuation adjustments with their explicit expression (5), we get

$$dI_t = I_t(r_t - \phi_t)dt - q_t \cdot S_t \mu_t dt + dM_t^I, \quad (12)$$

where we can see the dependency on the cost of carry  $\mu_t$  and on the allocation strategy.

## 2.2 The volatility targeting constraint

In a typical TVS, the fund manager selects a risky-asset portfolio with a specific time-dependent allocation strategy expressed by means of the vector of relative weights  $\alpha_t$ , along with a risk-free asset, which we can identify with the bank account  $B_t$ . In the following, for sake of simplicity in exposition, we consider only total-return securities, namely we set  $\pi_t = 0$ , which means that the security is not paying dividends. Thus we can write Equation (10) as given by

$$\frac{dI_t}{I_t} = \omega_t \alpha_t \cdot \frac{dS_t}{S_t} + (1 - \omega_t \alpha_t \cdot \mathbf{1}) \frac{dB_t}{B_t} - \phi_t dt, \quad (13)$$

where  $\mathbf{1}$  is a  $n$ -dimensional vector of ones and  $\omega_t \in [0, 1]$  is determined so that the strategy log-normal volatility is kept constant, namely

$$\omega_t : \quad \text{Var}_t[dI_t] = \bar{\sigma}^2 I_t^2 dt, \quad (14)$$

where  $\bar{\sigma}$  is the target volatility value. In practice, this means that the fund manager will select a risky-portfolio choosing  $\alpha_t$  equities from the universe where she can trade, and after that, her choices will be scaled by the automatic target volatility algorithm  $\omega_t$  (14).<sup>3</sup>

<sup>3</sup> We recall that the universe of assets where the manager can trade and the value of  $\bar{\sigma}$  are written in the contract.

It is possible to derive the expression for  $\omega_t$  assuming a generic continuous semi-martingales dynamics under the risk-neutral measure for the underlying securities, so that Equation (8) can be written as

$$\frac{dS_t^i}{S_t^i} = (r_t - \mu_t^i) dt + \nu_t^i \cdot dW_t, \quad (15)$$

where  $\nu_t$  is an adapted matrix process ensuring the existence of a solution for the stochastic differential equation and  $W_t$  is a  $n$ -dimensional vector of Brownian motions under  $\mathbb{Q}$ . Under these assumptions we can derive an expression for  $\omega_t$ , and we get<sup>4</sup>

$$\omega_t = \frac{\bar{\sigma}}{\|\alpha_t \cdot \nu_t\|}. \quad (16)$$

Hence, putting this last result in the dynamics of  $I_t$  we obtain

$$\frac{dI_t}{I_t} = \left( r_t - \phi_t - \frac{\bar{\sigma}\alpha_t}{\|\alpha_t \cdot \nu_t\|} \cdot \mu_t \right) dt + \frac{\bar{\sigma}\alpha_t}{\|\alpha_t \cdot \nu_t\|} \cdot \nu_t \cdot dW_t, \quad (17)$$

where we can see, as expected, that the strategy grows at the risk-free rate but for the valuation adjustments, given by the hedging costs expressed by the term proportional to  $\mu_t$  in (17), and the fees, given by the term  $\phi_t$  in (17).

We highlight that, to derive the dynamics expressed in Equation (17), we have not made any assumptions on the risky allocation strategy; thus all this argument is valid for any constraints on the process  $\alpha_t$  as, for instance, the case of contracts in which the fund manager is restricted to holding only long positions meaning that  $\alpha_t^i \geq 0 \forall i = 1, \dots, n$ , or situations where the combined total of long and short trades must be capped at a specific contractual amount.

### 3 Derivative pricing: target volatility options

In this section, we analyze TVO contracts, namely contracts linked to the TVS described previously. In particular, we will focus on European style options, and we will show that under appropriate assumptions it is possible to find a closed form solution for the optimal allocation strategy which maximizes the contract price.

In a general framework, a derivative contract on the TVS with maturity  $T$  can be defined as

$$V_0 := \sup_{\alpha} \mathbb{E}_0 \left[ \int_0^T D(0, u; \zeta_u) d\pi_u(\alpha) \right], \quad (18)$$

where  $D(0, T; \zeta_t)$  is the discount factor with rate  $\zeta_t$ , inclusive of the derivative valuation adjustments, and  $\pi_t$  is the cumulative coupon process paid by the derivative, and it depends on the allocation strategy since in turn the TVS depends on it via the valuation adjustments. We take the supremum over the

<sup>4</sup> In all formulae the norm for a vector  $a$  is defined as  $\|a\| := \sqrt{a \cdot a}$ .

strategies since we do not have any information on the future activity of the fund manager and we wish to calculate the worst-case scenario for the option seller, as discussed in the introduction of Section 2.

### 3.1 European options

If the derivative contract depends only on the marginal distribution of  $I_T$  at maturity (i.e., a European payoff), we are able to simplify the pricing problem. We consider the following pricing problem

$$V_0 := \sup_{\alpha} \mathbb{E}_0 [D(0, T; \zeta) \Phi(I_T(\alpha))] , \quad (19)$$

where  $\Phi$  is the payoff function of the derivative.

We start by introducing the Markovian projection of the dynamics followed by  $I_t$ . We name it  $I_t^{\text{MP}}$ , and we get by applying the Gyöngy Lemma [14]

$$\frac{dI_t^{\text{MP}}}{I_t^{\text{MP}}} := (r_t - \ell_{\alpha}(t, I_t^{\text{MP}})) dt + \bar{\sigma} dW_t^{\text{MP}} \quad \text{with} \quad I_0^{\text{MP}} = I_0 , \quad (20)$$

where the local drift is defined as

$$\ell_{\alpha}(t, K) := \bar{\sigma} \mathbb{E}_0 \left[ \frac{\mu_t \cdot \alpha_t}{\|\alpha_t \cdot \nu_t\|} \middle| I_t = K \right] , \quad (21)$$

and  $W_t^{\text{MP}}$  is a Brownian motion under the risk-neutral measure  $\mathbb{Q}$ . Notice that the diffusion coefficient collapses to the target volatility value  $\bar{\sigma}$ . Since European payoffs depend only on the marginal distribution at maturity, they can be calculated only by means of the Markovian projection  $I_t^{\text{MP}}$ , namely

$$V_0 := \sup_{\alpha} \mathbb{E}_0 [D(0, T; \zeta) \Phi(I_T^{\text{MP}}(\alpha))] . \quad (22)$$

Hence, we have our first result valid only if valuation adjustments can be neglected:

**Proposition 1.** *A European TVO can be calculated by assuming any allocation in the underlying risky basket if all the underlying securities grow under the risk-neutral measure at the risk-free rate without any valuation adjustment, namely if we can write  $\mu_t = 0$ .*

### 3.2 Stochastic optimal control problem

In presence of valuation adjustments, we need to solve the full optimization problem. We discretize the optimal strategy  $\alpha_t$  as<sup>5</sup>

$$\alpha_t := \sum_k \mathbf{1}_{\{t \in [T_{k-1}, T_k)\}} \alpha_{T_{k-1}} , \quad (23)$$

<sup>5</sup> We use the symbol  $\mathbf{1}_A$  for the indicator function of a subset  $A$ .

according to a time grid  $\mathcal{T} := \{T_0, \dots, T_k, \dots, T_m\}$  with  $T_0 := t$  the pricing date and  $T_m := T$  the maturity of the option. Therefore we can apply the dynamic programming principle to express the optimal  $\alpha_t$  at time  $T_{k-1}$  as

$$\alpha_{T_{k-1}} := \arg \max_{\alpha} \left\{ \mathbb{E}_{T_{k-1}} \left[ D(T_{k-1}, T_k) V_{T_k}(X_{T_k}, I_{T_k}(\alpha)) \mid X_{T_{k-1}}, I_{T_{k-1}} \right] \right\}, \quad (24)$$

where  $V_{T_k}$  is the option value at time  $T_k$  and  $X$  is any Markovian state such that the drift and the diffusion coefficient of  $I_t$  are a function of  $(X_t, I_t, \alpha_t)$  and we indicated with  $I_{T_k}(\alpha)$  the value of the strategy at time  $T_k$  for a given choice of the weights  $\alpha$  at time  $T_{k-1}$ . We calculate the strategy value for the optimal weights  $\alpha_{T_{k-1}}$ , namely  $I_{T_k}(\alpha_{T_{k-1}})$  by a suitable discretization of (17) starting from  $X_{T_{k-1}}$  and  $I_{T_{k-1}}$ . Thus, once collected the elements  $\{X_{T_k}, I_{T_k}(\alpha_{T_{k-1}})\}$ , the derivative price is given by:

$$\mathbb{E}_{T_{k-1}} \left[ D(T_{k-1}, T_k) V_{T_k}(X_{T_k}, I_{T_k}(\alpha_{T_{k-1}})) \mid X_{T_{k-1}}, I_{T_{k-1}} \right] = V_{T_{k-1}}(X_{T_{k-1}}, I_{T_{k-1}}) = \quad (25)$$

while the iteration starts from maturity date where the boundary condition is set equal to the payoff function:

$$V_{T_m} = \Phi(I_{T_m}). \quad (26)$$

### 3.3 Black and Scholes scenario

In the time-dependent Black and Scholes model with deterministic rates, we can work with empty  $X_t$ , since in this case the portfolio dynamics (17) is Markovian, leading to an optimal strategy  $\alpha_t^*$  which depends in principle only on  $I_t$ . As a consequence, the local drift defined in Equation (21) can be written as

$$\ell_{\alpha}(t, K) = \bar{\sigma} \frac{\mu(t) \cdot \alpha(t, K)}{\|\alpha(t, K) \cdot \nu(t)\|}, \quad (27)$$

so that the optimization problem can be solved by looking only at the Markovian projection without simulating all the Brownian motions  $W_t$ . Notice that we are indicating the dependency on time in parenthesis to highlight that in this formula all the quantities are deterministic functions of time.

A direct consequence is the following proposition, which is relevant for plain vanilla options on TVS.

**Proposition 2.** *When the underlying securities follow a Black and Scholes model with deterministic rates, the optimal strategy for a non-decreasing European payoff consists in minimizing the local drift function, independently of the current state  $I_t$*

$$\alpha^*(t) := \arg \min_{\alpha} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|}. \quad (28)$$

Analogously, the optimal strategy for a non-increasing European payoff consists in maximizing the local drift function:

$$\alpha^*(t) := \arg \max_{\alpha} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|}. \quad (29)$$

The absence of stochastic elements in Equation (28) makes the optimal strategy known *a priori* with no numerical simulation needed to solve the control problem; in fact, the functions  $\mu(t)$  and  $\nu(t)$  can be directly derived from the market quotes of the risky assets and of its derivative contracts, and then we can apply Equation (28)  $\forall t \in \mathcal{T}$  to retrieve the deterministic optimal strategy. Once  $\alpha^*$  is known, then one can price the payoff by the following BS formula

$$V_0^{\text{BS}} = BS(F^{\text{TVS}}(0, T; \alpha^*), K, T, \bar{\sigma}, D(0, T; \zeta)) , \quad (30)$$

where  $F^{\text{TVS}}(t, T; \alpha)$  is the TVS forward curve defined by

$$F^{\text{TVS}}(t, T; \alpha) = I_t \exp \left[ \int_t^T (r(u) - \ell_\alpha(u)) du \right] , \quad (31)$$

while  $BS(F, K, T, \sigma, D)$  is the standard BS formula for a European option with forward curve  $F$ , strike  $K$ , time to maturity  $T$ , volatility  $\sigma$  and discount factor  $D$ .

In Figure 1 we provide a comparison among the option price obtained with the optimal strategy (BS\*) of Equation (28) and those with the below expert-based single-asset strategies ( $S_A$ ,  $S_B$ ,  $S_C$ ). Here we consider the case of an at-the-money call option with spot  $I_0 = 1$  EUR, maturity  $T = 5$  yr, target volatility  $\bar{\sigma} = 5\%$ , and a nonnegative constraint on  $\alpha$ . The intuitive strategies are

- $S_A$ : invest all in the asset with the maximum forward curve at maturity; this strategy consists of selecting a static portfolio that is never rebalanced, such that the TVO moneyness is maximized;
- $S_B$ :  $\forall t \in \mathcal{T}$  invest all in the asset with minimum  $\mu(t)$ ; in this case we consider a fund manager who rebalances the weights  $\alpha(t)$  over the life of the portfolio by selecting the asset with the highest expected return at a given point in time;
- $S_C$ :  $\forall t \in \mathcal{T}$  invest all in the asset with minimum  $\mu(t)/\|\nu(t)\|$ ; the strategy is akin to  $S_B$ , with the exception being that the manager endeavors to maximize the expected return that has been adjusted to account for the target volatility mechanism.

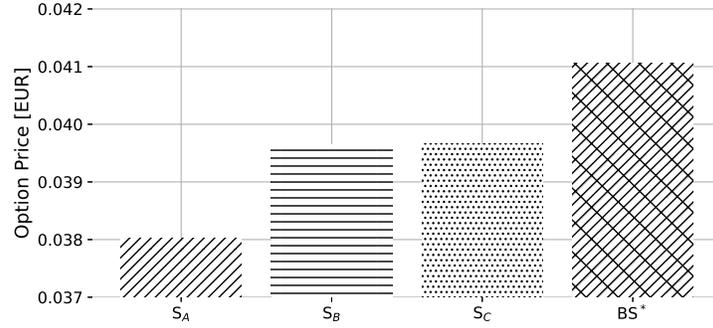
As the reader can observe, our strategy outperforms any other expert-based one a practitioner might adopt to face the control problem.

In the next subsection we derive the analytical solution of the problem (28) under the assumption of unconstrained allocation strategy  $\alpha_t$ .

**Unconstrained allocation strategy: closed form solution** In absence of constraints on the allocation strategy, we are able to derive a closed form solution to the BS problem (28).

**Lemma 1.** *Let  $\mu, \alpha \in \mathbb{R}^n$ ,  $\nu \in \mathbb{R}^{n \times n}$  be a full rank matrix and  $\Sigma := \nu\nu^\top$ . Then the closed solution of the optimization problem (28) is*

$$\alpha^* = - \frac{\Sigma^{-1} \cdot \mu}{\|(\Sigma^{-1} \cdot \mu) \cdot \nu\|} . \quad (32)$$



**Fig. 1.** Comparison of plain vanilla prices on the TVS by adopting different allocation strategies: the optimal Black and Scholes solution ( $BS^*$ ) of Equation (28) and three expert-based single-asset strategies ( $S_A, S_B, S_C$ )

*Proof.* Since the argument of the minimum (28) is zero-homogeneous, then we can rewrite the problem as

$$\begin{aligned} & \text{minimize } \alpha \cdot \mu \\ & \text{subject to } \|\alpha \cdot \nu\|^2 = 1. \end{aligned} \quad (33)$$

By setting the Lagrangian function associated with the problem

$$\mathcal{L}(\alpha, \lambda) = \alpha \cdot \mu - \lambda (\|\alpha \cdot \nu\|^2 - 1), \quad (34)$$

we obtain the first order conditions

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \alpha} = \mu - 2\lambda \Sigma \cdot \alpha = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = \|\alpha \cdot \nu\|^2 - 1 = 0, \end{cases} \quad (35)$$

Then, by applying simple algebra, we obtain the analytical form for the free optimal strategy

$$\alpha^* = \pm \frac{\Sigma^{-1} \cdot \mu}{\|(\Sigma^{-1} \cdot \mu) \cdot \nu\|}. \quad (36)$$

We take the minus sign to get the minimum value of the TVS local drift while the plus sign for the maximum one (put payoff).  $\square$

In the following subsection we prove that an analytical solution of (28) exists also assuming a nonnegative constrained allocation strategy. Under this assumption, we show that the optimal strategy consists in investing in a single asset. We name this strategy the bang-bang solution.

**Active asset (or bang-bang) solution** A closed form solution to the minimization of the local drift correction (28) can also be derived in the common case that all costs of carry  $\mu$  are nonnegative, so they represent true hedging costs and not benefits, and the only constraint on portfolio weights is nonnegativity, which would mean a long-only strategy by the fund manager.

**Lemma 2.** *Let  $\mu \in \mathbb{R}_+^n$  be a vector with nonnegative components,  $\nu \in \mathbb{R}^{n \times n}$  be a full rank matrix, and  $\Sigma = \nu\nu^\top$ . Then*

$$\inf_{\alpha \in \mathbb{R}_+^n \setminus \{0\}} \frac{\alpha \cdot \mu}{\|\alpha \cdot \nu\|} = \min_{i \leq n} \frac{\mu_i}{\sqrt{\Sigma_{ii}}} ; \quad (37)$$

if  $\bar{i}$  is the index which realizes the min, then the infimum is realized by a vector concentrated on the  $\bar{i}$  component:  $\alpha_i = \delta_{i\bar{i}}$ .

*Proof.* Let us first consider the case in which  $\mu = \mathbb{1}$ . Since the argument of the infimum is zero-homogeneous, normalizing by  $\alpha \cdot \mathbb{1} > 0$  we can restrict to the affine hyperspace  $\{\alpha \cdot \mathbb{1} = 1\}$ , where the minimization (37) reduces to the maximization of its denominator: the required infimum will be the square root of the reciprocal of

$$\sup \{ \|\alpha \cdot \nu\|^2 \mid \alpha \in \mathbb{R}_+^n, \alpha \cdot \mathbb{1} = 1 \} . \quad (38)$$

Now we can note that  $\Sigma$  is positive definite, hence  $\Sigma_{ij} < \sqrt{\Sigma_{ii}\Sigma_{jj}} \leq \Sigma_{\bar{i}\bar{i}}$ , which implies

$$\|\alpha \cdot \nu\|^2 = \sum_{i,j=1}^n \alpha_i \alpha_j \Sigma_{ij} \leq \sum_{i,j=1}^n \alpha_i \alpha_j \Sigma_{\bar{i}\bar{i}} = \Sigma_{\bar{i}\bar{i}} , \quad (39)$$

because  $\sum_i \alpha_i = 1$ . Since we trivially have equality for  $\alpha_i = \delta_{i\bar{i}}$ , this concludes the proof of the case  $\mu = \mathbb{1}$ .

Next, let us consider the case in which all components of  $\mu$  are strictly positive, and define  $M$  as the diagonal matrix with diagonal  $\mu$ . Then we can rewrite the infimum as a function of  $\beta = M\alpha$ :

$$\inf_{\beta \in \mathbb{R}_+^n \setminus \{0\}} \frac{\beta \cdot \mathbb{1}}{\|\beta \cdot M^{-1}\nu\|} , \quad (40)$$

which by the first part of the proof equals

$$\min_{i \leq n} \frac{1}{\sqrt{\tilde{\Sigma}_{ii}}} = \min_{i \leq n} \frac{\mu_i}{\sqrt{\Sigma_{ii}}} , \quad \tilde{\Sigma} := M^{-1}\nu\nu^\top M^{-1} = M^{-1}\Sigma M^{-1} . \quad (41)$$

Finally, let us consider the general case in which  $\mu$  may have some components equal to zero. For an arbitrary  $\epsilon \geq 0$  let us define

$$f_\epsilon(\alpha) = \frac{\alpha \cdot (\mu + \epsilon)}{\|\alpha \cdot \nu\|} . \quad (42)$$

One can easily note that as  $\epsilon \rightarrow 0$ ,  $f_\epsilon$  tends to  $f_0$  uniformly on the compact set  $\{\alpha \in \mathbb{R}_+^n \mid \alpha \cdot \mathbb{1} = 1\}$ , so that the minimum converges to the minimum on that

set. Since we know by homogeneity that the minimum on  $\{\alpha \in \mathbb{R}_+^n \mid \alpha \cdot \mathbb{1} = 1\}$  equals the minimum on  $\mathbb{R}_+^n \setminus \{0\}$ , we conclude

$$\inf_{\alpha \in \mathbb{R}_+^n \setminus \{0\}} f_0(\alpha) = \lim_{\epsilon \rightarrow 0^+} \inf_{\alpha \in \mathbb{R}_+^n \setminus \{0\}} f_\epsilon(\alpha) = \lim_{\epsilon \rightarrow 0^+} \min_{i \leq n} \frac{\mu_i + \epsilon}{\sqrt{\Sigma_{ii}}} = \min_{i \leq n} \frac{\mu_i}{\sqrt{\Sigma_{ii}}} . \quad (43)$$

□

### 3.4 Hamilton-Jacobi-Bellman equation for target volatility options

In this section, we provide to the reader a formal description of the dynamic problem associated with options on target volatility strategies by writing the Hamilton-Jacobi-Bellman (HJB) equation for the derivative price. We prove that from this equation we can recover the same closed formula (28) derived above from the Gyöngy Lemma for the time-dependent BS model.

In full generality, we assume that the time evolution of the Markovian factors governing the problem dynamics is given by a stochastic multidimensional process in  $\mathbb{R}^n$ ,  $X_t$ , which is the unique strong solution to the following Itô stochastic differential equation

$$dX_t = M(X_t)dt + \Sigma(X_t) \cdot dW_t , \quad (44)$$

where  $W_t$  is a  $n$ -dimensional Wiener process with independent components. We point out to the reader that with this notation we are including general dynamics models like those with stochastic drift  $M(X_t)$  and stochastic diffusive  $\Sigma(X_t)$  coefficients.

Let be the dynamics of the securities  $S_t$  a generic function of the Markovian factors, namely

$$S_t := f(X_t) . \quad (45)$$

In this framework, the TVS price process dynamics is given by the stochastic differential equation

$$\frac{dI_t}{I_t} = \left( r(X_t) - \frac{\bar{\sigma} \alpha_t}{\|\alpha_t \cdot \nu(X_t)\|} \cdot \mu(X_t) \right) dt + \frac{\bar{\sigma} \alpha_t}{\|\alpha_t \cdot \nu(X_t)\|} \cdot \nu(X_t) \cdot dW_t , \quad (46)$$

where the expression of  $\mu(X_t)$  and  $\nu(X_t)$  can be recovered by applying the Itô formula to Equation (45).

Given  $X := X_t$  and  $I := I_t$ , we can write the HJB equation for  $V := V(t, I, X)$  as follows

$$\begin{aligned} \frac{\partial V}{\partial t} + \max_{\alpha} \left\{ \left( r(X) - \bar{\sigma} \frac{\alpha \cdot \mu(X)}{\|\alpha \cdot \nu(X)\|} \right) I \frac{\partial V}{\partial I} + (\nabla_X V) \cdot M(X) + \frac{1}{2} \bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} \right. \\ \left. + \frac{1}{2} \text{Tr} (\Sigma(X)^\top (H_X V) \Sigma(X)) + (\nabla_{X,I} V) \cdot \Sigma(X) \cdot \left( I \bar{\sigma} \frac{\alpha \cdot \nu(X)}{\|\alpha \cdot \nu(X)\|} \right) \right\} = 0 , \end{aligned} \quad (47)$$

where  $\text{Tr}(A)$  is the trace operator of  $A$ ,  $\nabla_X V$  the gradient of  $V$  w.r.t.  $X$ ,  $H_X V$  the Hessian matrix of  $V$  w.r.t.  $X$  and  $\nabla_{X,I} V$  is the vector defined by:

$$\nabla_{X,I} V := \left( \frac{\partial^2 V}{\partial X^1 \partial I}, \dots, \frac{\partial^2 V}{\partial X^n \partial I} \right)^\top. \quad (48)$$

We take out from the maximum operator all the elements that do not depend on the risky allocation strategy  $\alpha$

$$\begin{aligned} \frac{\partial V}{\partial t} + r(X)I \frac{\partial V}{\partial I} + (\nabla_X V) \cdot M(X) + \frac{1}{2} \bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} + \frac{1}{2} \text{Tr}(\Sigma(X)^\top (H_X V) \Sigma(X)) \\ + \bar{\sigma} I \max_{\alpha} \left\{ -\frac{\partial V}{\partial I} \frac{\alpha \cdot \mu(X)}{\|\alpha \cdot \nu(X)\|} + (\nabla_{X,I} V) \cdot \Sigma(X) \cdot \left( \frac{\alpha \cdot \nu(X)}{\|\alpha \cdot \nu(X)\|} \right) \right\} = 0. \end{aligned} \quad (49)$$

Equation (49) is the HJB equation describing the TVS dynamic problem for a generic dynamics of the risky securities underlying the portfolio.

If we assume a time-dependent BS dynamics for the risky equities ( $\mu_t$ ,  $r_t$  and  $\nu_t$  deterministic), then  $V = V(t, I)$  and all the derivatives w.r.t.  $X$  are zero. Therefore the reduced HJB equation is

$$\frac{\partial V}{\partial t} + r(t)I \frac{\partial V}{\partial I} + \frac{1}{2} \bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} + \bar{\sigma} I \max_{\alpha} \left\{ -\frac{\partial V}{\partial I} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|} \right\} = 0; \quad (50)$$

if the payoff is non-decreasing in  $I$ , by homogeneity of the stochastic differential equation we get that  $V$  is non-decreasing. Thus the solution is given by

$$\alpha^*(t) = \arg \min_{\alpha} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|}, \quad (51)$$

which is the same result expressed in Equation (28). On the other hand, if the payoff is non-increasing in  $I$  then the solution will be the arg max.

Conversely, if we deal with a dynamic model for which the derivative contract  $V$  depends on  $X$  then the volatility versor term, namely the second one, in the max operator of Equation (49) is no longer zero and thus one must solve the entire control problem numerically. Indeed, one can check by direct substitution that the natural generalization of (51)

$$\alpha^*(X_t) = \arg \min_{\alpha} \frac{\alpha \cdot \mu(X_t)}{\|\alpha \cdot \nu(X_t)\|} \quad (52)$$

fails to satisfy (49).

In what follows we tackle the non-trivial case of a local volatility model for the  $S_t$ -dynamics, such that  $\nu_t = \nu(t, S_t)$ . We have chosen the LV model since it is well known in financial literature and among practitioners [11].

## 4 Reinforcement learning

As we have discussed in the previous sections, the analytical solution described in Proposition 2 crucially relies on the assumption that the underlyings evolve

according to a Black and Scholes model, and on a monotonicity assumption on the payoff. Therefore, one must resort to numerical approaches to solve the stochastic control problem related to the TVS in the case of general payoffs or risky securities dynamics. The standard approach could be to use classical techniques based on backward recursion (24)-(25) such as American Monte Carlo [20]. However, their performances degrade exponentially as the dimension  $n$  of the problem increases, making prohibitively costly finding the solution to the problem. In our contribution, we adopt a novel technique which is free from the curse of dimensionality and is gaining popularity in many scientific branches for solving stochastic optimal control problems: Reinforcement Learning (RL) [27].

RL is a branch of Machine Learning which allows an artificial agent to interact with an environment through actions and observations in order to maximize some notion of cumulative reward called value function. In RL, the agent is not told which actions to take but instead, it has to discover by trial and error which are the behaviors yielding the highest reward. This is obtained by updating the agent policy  $\pi$  which is a mapping from the environment states to the set of actions. Thus RL is independent of pre-collected data as opposed to other Machine Learning techniques, that have to be fed with a pre-existing dataset to learn from. Because of its nature, RL has been successful in quantitative finance for solving control problems; among the most important RL applications in this field, we refer to [4] as the pioneers in studying self-taught reinforcement trading problems, while to [15] and [19] for hedging derivatives with RL under market frictions.

In our work we adopt two learning strategies to compare their performances in terms of overall reward: a plain direct policy search algorithm [27] and the state of the art proximal policy optimization (PPO) developed in [25] and [24].

The first method is well-known in RL literature: it is based on the direct update of the agent policy via a gradient-based optimization of the value function. We will go into details in Section 4.1.

On the other hand, the PPO is a high-level actor-critic algorithm well-suited for continuous control problems. It collects a small batch of experiences interacting with the environment to update its decision-making policy. From those interactions with the environment, PPO is able to compute the expected reward and the value function. We will not provide a complete description of this sophisticated learning strategy; for more details, we refer to the authors' papers. In our work, we adopt the PPO implementation found in OPENAI BASELINES.<sup>6</sup>

In the following sections, we describe the way we have formalized the TVS problem in the Reinforcement Learning framework.

#### 4.1 Direct policy search approach

In RL, the learning phase takes place within a set of so-called episodes, which refer to a single run or instance of interaction between the RL agent and an

---

<sup>6</sup> <https://github.com/openai/baselines>.

environment, where the agent takes a sequence of actions to achieve a specific goal.

For our problem, we consider an episode  $\tau$  of length  $m+1$  that unfolds over a discrete time-grid of fixing times expressed in year fractions  $\mathcal{T} := \{T_0, \dots, T_k, \dots, T_m\}$  with  $T_0 := 0$  and  $T_m := T$  maturity of the option. At a given episodic time  $T_k$  the RL agent interacts with the environment: it receives a representation of the environment called state  $s_k$  and on the basis of that it selects an action  $a_k$  sampling from the current policy  $\pi^{\theta_j}$ . Here with  $\theta_j$  we refer to the set of parameters through which we parameterize the agent policy at the  $j$ -th algorithm iteration. In our case of study, the RL agent, representing the fund manager, has the right to select the composition of the risky asset portfolio, so that the policy is the allocation strategy  $\alpha$  introduced in Equation (13):

$$a_k = \alpha_{T_k} \quad \forall T_k \in \mathcal{T}, \quad a_k \in \mathcal{A} \subset \mathbb{R}^n. \quad (53)$$

Since the value function of the problem depends on the Markovian state  $X_t$ , the portfolio level  $I_t$  and time  $t$ , our natural choice for the observation state is the following block

$$s_k := [X_{T_k}, I_{T_k}, T_k] \quad \forall T_k \in \mathcal{T}, \quad s_k \in \mathcal{S} \subset \mathbb{R}^{n+2}. \quad (54)$$

In this way, the state contains all the information needed by the agent to select the optimal action, leading to the maximum plain vanilla TVO price.

Once the agent has selected the action  $a_k$  sampled from the current policy, it receives at the next time  $T_{k+1}$  a reward  $r_{k+1}$  generated by the environment. In this algorithm we have defined the reward function as follows

$$r_{k+1} := \begin{cases} D(0, T; \zeta)(I_T(\alpha(\theta_j)) - K)^+ & \text{if } T_{k+1} = T \\ 0 & \text{otherwise} \end{cases}. \quad (55)$$

Therefore, during the whole episode, the agent receives a nil reward except at maturity when the reward coincides with the option intrinsic value. This choice may seem too daring because the agent receives a real feedback of its actions only at the end of the episode, which could result in a slower learning. However, if the agent has learnt the optimal policy  $\pi^* = \pi^{\theta^*}$ , the average cumulative reward per episode will coincide with the optimal TVO price.

In this algorithm, we parameterize the agent policy with a feed forward neural network (FFNN), such that  $\theta_j$  coincides with the hidden weights,  $s_k$  with the input neurons, and  $a_k$  with the output ones. In this way, we are dealing with a deterministic policy, that maps each state directly to a single action without any additional randomness or probability distribution; the map functional form is given by the neural network. The parameters update is performed as follows: the agent collects a finite batch of experiences interacting with the environment in a set of episodes  $\tau$ , then the loss-value function is evaluated as

$$L(\theta_j) = \hat{\mathbb{E}} \left[ \sum_{k=0}^{m-1} r_{k+1} | \theta_j \right], \quad (56)$$

where the expectation  $\hat{\mathbb{E}}[\dots]$  indicates the empirical average on the batch. Then the parameters are updated by plugging the policy into a stochastic gradient ascent algorithm [27] with the aim of maximizing Equation (56).

Once the training phase is ended and the agent has selected the optimal policy, we can run a Monte Carlo (MC) simulation with never seen scenarios to price the optimal target volatility option and test if the algorithm has not overfitted the data.

## 4.2 Proximal policy optimization approach

As for the direct policy search approach, we model the pricing problem considering an episode that takes place on the time-grid  $\mathcal{T} := \{T_0, \dots, T_k, \dots, T_m\}$  with  $T_0 := 0$  and  $T_m := T$ . Again we choose as observation state the block defined in Equation (54) and the agent policy coincides with the risky allocation strategy  $\alpha$  (53). As the PPO algorithm is more sophisticated than the previous one, we have decided to test its performance by means of two different reward functions. The former is the same as defined by Equation (55), while the latter is given by

$$r_{k+1} := \gamma^k [V_{\text{BS}}(T_{k+1}) - V_{\text{BS}}(T_k)], \quad (57)$$

where  $\gamma \in [0, 1]$  is an hyper-parameter of the PPO, while  $V_{\text{BS}}(T_k)$  is a proxy of the residual option price defined by

$$\begin{aligned} V_{\text{BS}}(T_k) &:= BS(F^{\text{TVS}}(T_k, T; \alpha_{\text{BS}}^*, K), K, T - T_k, \bar{\sigma}, D(T_k, T; \zeta)), \\ V_{\text{BS}}(T_0) &= 0, \end{aligned} \quad (58)$$

with  $\alpha_{\text{BS}}^*$  the BS optimal strategy (28) evaluated at the state  $s_k$ . In this form the agent actions are hidden inside the term  $I_t$  used to compute the TVS forward curve  $F^{\text{TVS}}$  defined by Equation (31). In this case, the reward function does not suffer of nil values for  $0 < T_k < T$ : the RL agent always gets a feedback from the environment for its choices. Notice that the hyper-parameter  $\gamma$  plays the role of a discount factor in the sense that, as  $\gamma$  approaches to zero, the RL agent will tend to maximize immediate rewards while neglecting possible larger rewards in the future. If we take the cumulative reward per episode and set the PPO parameters<sup>7</sup>  $\gamma = \lambda = 1$  we obtain

$$R(\tau) = \sum_{k=0}^{m-1} r_{k+1} \stackrel{\gamma=\lambda=1}{=} \sum_{k=0}^{m-1} [V_{\text{BS}}(T_{k+1}) - V_{\text{BS}}(T_k)] = V_{\text{BS}}(T) = (I_T - K)^+, \quad (59)$$

which is equal to the intrinsic value of the option. This result does not depend on the definition of  $V_{\text{BS}} \forall T_k < T$ , but we conjecture that the closer  $V_{\text{BS}}$  is to the value function, the easier the agent is in learning.

Thus one can train the agent choosing the optimal value for  $\gamma, \lambda \in [0, 1]$ , and then run, as test phase, a Monte Carlo simulation with  $\gamma = \lambda = 1$  and the

<sup>7</sup> We refer to [24] for a more detailed description for the generalized advantage estimation parameter  $\lambda$ .

optimized  $\theta$  fixed, where, if the agent has learnt  $\pi^*$ , the average of  $R(\tau)$  along different episodes will match the optimal undiscounted price of the derivative contract on the TVS.

In the OPENAI BASELINES implementation of the PPO, the agent policy is parameterized again by a neural network; as for the previous method we have chosen an FFNN. More precisely, training is based on a stochastic policy, i.e. the action is drawn from a probability law, and the neural network is used to determine its parameters. In settings like ours where the action has a continuous domain, such law is a multivariate diagonal Gaussian distribution where the mean  $\mu^\theta(s_k)$  is the output vector of the FFNN and the log-standard deviation  $\log \sigma$  is an additional trainable parameter:

$$\pi^\theta(s_k) \sim \mathcal{N}(\mu^\theta(s_k), e^{\log \sigma}) . \quad (60)$$

As one can observe,  $\log \sigma$  is state-independent, but it is reduced as the number of the PPO update iterations increases. The idea is that the log-standard deviation will be higher at the beginning of the training phase in order to guarantee a good exploration of the action space while it will be lower at the end to avoid too much noise in the proximity of the optimal policy. Eventually at the end of training, the standard deviation is put to zero, so that the action becomes deterministically equal to the output of the FFNN.

The fact that the PPO implementation exploits a stochastic policy ensures us a better exploration of the action space than with the previous approach. Moreover, the algorithm tries to learn an approximator of the on-policy value function as control variate for the training phase. This approximator is an FFNN with the same architecture as the one for the policy.

## 5 Numerical investigations

Here we present the numerical results obtained with our proposed methods. We focus our analysis on a European call option on a TVS with the following product details

$$I_0 = K = 1 \text{ EUR} , \quad T = 2 \text{ yr} , \quad \bar{\sigma} = 5\% ,$$

where  $I_0$  is the starting value of the TVS-based portfolio. Moreover, without loss of generality, we consider the case of an unconstrained allocation strategy  $\alpha$ . The extension to the constrained case is easy.

It is our aim to investigate the control problem under non-trivial dynamics like the local volatility model where the volatility of the risky asset is also a function of the state. By looking at the HJB equation (49), we expect that the volatility versor will play a role in finding the optimal solution, giving rise to a non-trivial strategy.

Although in Section 3.3 we have proved that under the Black and Scholes model the Equation (28) solves the control problem, we want to take advantage of this *a priori* solution as a benchmark to gather evidence on the robustness of our RL approach and to check if our analytical result is correct. Moreover, we

use the BS model as numerical laboratory to perform fine-tuning tests for the RL algorithms hyper-parameters and analyze how they impact the final results and performances.

We recall that in both the algorithms we parameterize the agent policy with an FFNN; thus this is completely characterized by the following hyper-parameters: number of hidden layers, number of neurons per hidden layer, and activation function per hidden layer. This is due to the fact that in this RL problem the number of the input neurons is equal to the state space dimension (54), while the number of the output ones coincides with the action space dimension (53). It is well known in the literature that neural networks give better performances in the training phase if the input data are well normalized [23], [26]. Thus we choose as state  $s_k$  the following normalized block

$$[\log(S_{T_k}/F(0, T_k)), I_{T_k}/I_0, T_k] \quad \forall T_k \in \mathcal{T}, \quad (61)$$

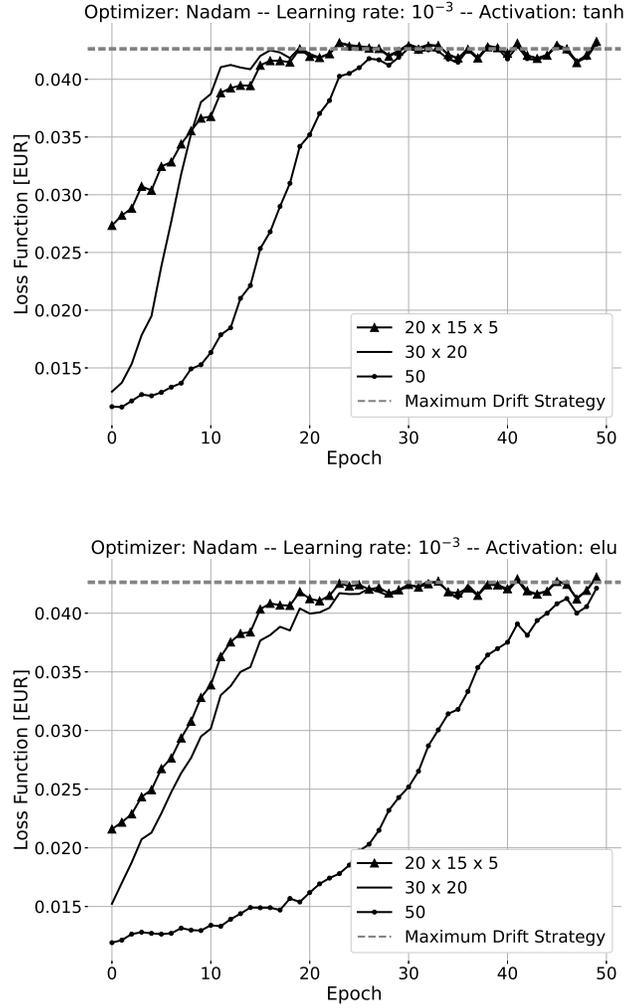
where  $F(t, T)$  is the forward curve vector of the risky assets from  $t$  to  $T$ . In Equation (61) we have chosen as Markovian state  $X_{T_k}$  the martingale term of the securities dynamics. In this way, we have that in the input block the variables have the same order of magnitude.

### 5.1 Black and Scholes: hyper-parameters fine tuning

We use the BS environment as a toy model to understand which parameters of the RL algorithm play key roles in the training and testing phase. Our first approach has been the direct policy one since it represents a natural way to tackle the problem: since our goal is to find the allocation strategy that maximizes the option price, we update the FFNN weights following the gradient direction of the loss function defined in Equation (56).

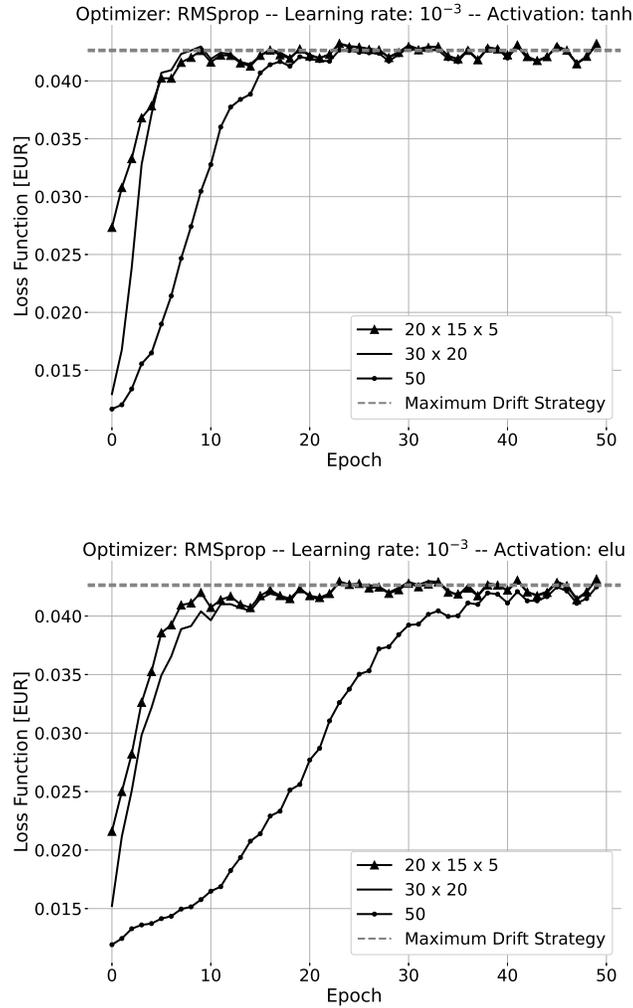
We try to investigate the following hyper-parameters: the FFNN architecture, in particular which feature between the depth and the width of the network is more important, the activation function, and the learning rate of the optimizer. Moreover, we compare the performances of two well-known optimizers in Machine Learning literature: Nadam [8] and RMSprop [16]. Since we deal with a free allocation strategy  $\alpha$  that can assume negative values, we have chosen among the wide variety of activation functions[12] the tanh and the elu. In this way, we can analyze the performance of a saturating activation function and a non-saturating one. Firstly we have performed a grid search on the learning rate starting value and we have found  $10^{-3}$  a good choice in terms of speed of learning and avoiding over-fitting. Note that to detect over-fitting one typically compares the performance on one or more training sets with that on corresponding validation sets. Since our environment is simulated, there is no need to explicitly set aside training data for such task, nor to perform  $K$ -fold validation: indeed we can always generate for the purpose a set of new scenarios which were not seen during optimization.

In Figures 2 and 3 we present the learning curves of our fine-tuning tests for RMSProp and Nadam respectively. The three lines of each plot display the



**Fig. 2.** Learning curves of the direct policy algorithm applied to the Black and Scholes scenario. The solid lines are the loss functions defined in Equation (56) in function of the number of training epochs. Each continuous line represents a different neural network architecture with tanh activation function (top) and elu activation function (bottom). The optimizer adopted is the Nadam with a learning rate of  $10^{-3}$ . The horizontal grey-dashed line is the conservative option price obtained by maximizing the TVS drift through Equation (36)

learning curves for different FFNN architectures: a one hidden layer network with 50 neurons, a 2 hidden layers with 30 and 20 neurons respectively, and a deeper



**Fig. 3.** Learning curves of the direct policy algorithm applied to the Black and Scholes scenario. The solid lines are the loss values defined in Equation (56) in function of the number of training epochs. Each continuous line represents a different neural network architecture with tanh activation function (top) and elu activation function (bottom). The optimizer adopted is the RMSprop with a learning rate of  $10^{-3}$ . The horizontal grey-dashed line is the conservative option price obtained by maximizing the TVS drift through Equation (36)

one with three layers with 20, 15, and 5 neurons. Each learning curve is compared with the optimal option price (grey-dashed line) that we have computed with

the closed form solution we have derived for the BS model in Section 3.3. We can see that all the learning curves converge to our expected price, providing a numerical finding of our theoretical result. More deeply, from the plots we observe that the RMSprop optimizer outperforms the Nadam. Moreover, the tanh activation function seems to be more preferable than the non-vanishing elu. However, more importantly, we have evidence that a deeper architecture of the neural network outperforms the shallow one. All the learning curves we have presented are the best-in-sample results, in terms of performance, of four runs with different random starting guesses for the hidden weights  $\theta$ . This procedure is necessary since the objective function is not convex.

We take the  $20 \times 15 \times 5$  network with tanh from RMSprop as the best optimized network, and we run a Monte Carlo simulation with  $10^6$  never-seen scenarios to check if the agent overfits the new data. We report the results in

**Table 1.** Comparison of TVO prices under Black and Scholes scenario: analytical solution price and direct policy reinforcement learning. The option parameters are:  $I_0 = K = 1$  [EUR],  $T = 2$  [yr] and  $\bar{\sigma} = 5\%$

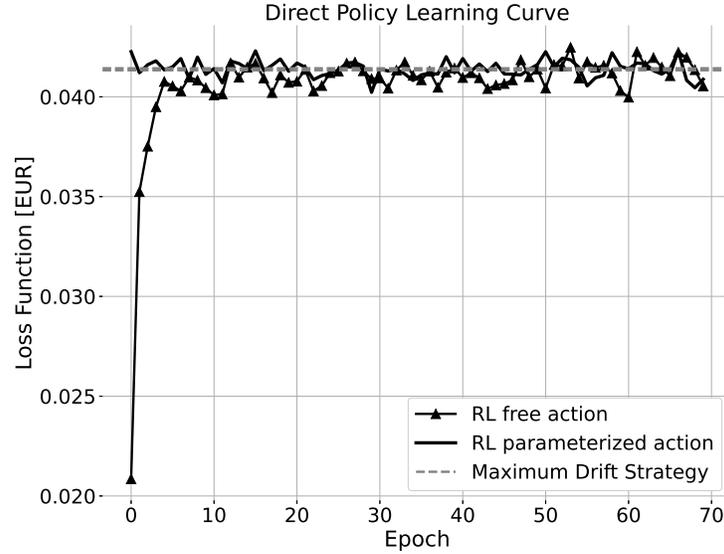
Pricing method	TVO price [EUR]
Analytical Solution	$4.2634 \times 10^{-2}$
Direct policy RL	$(4.2624 \pm 0.005) \times 10^{-2}$

Table 1: the RL price is compatible with the closed formula price. Thus the RL agent did not overfit the data during the training phase.

We will take advantage of those fine-tuning results to tackle the local volatility problem.

## 5.2 Local volatility dynamics

In this section, we study the TVS control problem assuming a local volatility model for the dynamics of the risky assets. Thus in this case we have a diffusive term in Equation (15) that is a deterministic function both of time and state, i.e. the spot price,  $\nu_t = \nu(t, S_t)$ . This additional dependency of volatility makes the problem of finding the optimal strategy non-trivial, as a closed form formula cannot be derived under this scenario; in fact if we consider the whole volatility smiles of the securities, then the second-order term in the HJB equation (49) does not elide and consequently the only way to derive  $\alpha^*$  is through a numerical method. The additional dependence on the state  $S_t$  under LV dynamics will cause the agent to consider also the  $X_t$  component value in the state block (54), unlike under the BS scenario where this information is redundant as the optimal strategy was function only of time (see Proposition 2). Here we consider the same market data ( $r_t$ ,  $\mu_t$  and  $\nu_t$ ) as in the Black and Scholes environment to study how the optimal solution changes with the dynamics model.



**Fig. 4.** Learning curves of the direct policy algorithm applied to the local volatility scenario. The solid line with triangles is the learning curve of an agent with free actions, while the solid one is the learning curve of an agent whose actions are parameterized with the maximum drift strategy. The two agents are parameterized by a  $20 \times 10 \times 5$  FFNN with tanh activation function. The loss function is computed according to Equation (56). The optimizer adopted is the RMSprop with a learning rate of  $10^{-3}$  for the solid line with triangles while  $10^{-4}$  for the solid one. The horizontal grey-dashed lines delimit the 99% confidence interval of the MC price obtained by maximizing path-wise the TVS drift through Equation (36)

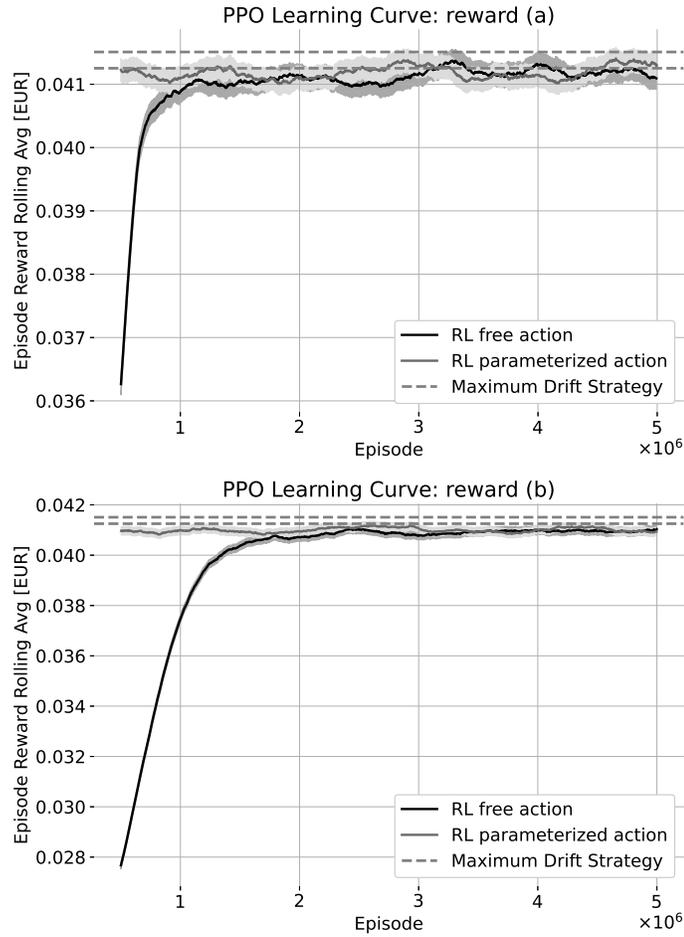
Our first way to tackle the problem is to exploit the results obtained in the BS case. Thus we train by direct policy algorithm a deep  $20 \times 10 \times 5$  neural network with tanh activation function and a RMSprop optimizer. We report in Figure 4 the corresponding learning curve (solid line with triangles). From this, we observe that the network has learnt some good policy since the curve grows as the number of training epochs increases until it saturates at a certain value. To try to measure the performance of the policy selected by the agent, we can build a naïf strategy called “baseline”. By looking at the Markovian projection in (20)-(21) and the first term in the HJB equation (49), we see that the control problem is similar to the BS one with an additional second order term. Thus a natural choice for the baseline is to maximize the TVS local drift by applying the path-wise the Black and Scholes solution of Equation (28). Since here we deal with a free allocation strategy, we can simply use our analytical result (36). In the same Figure 4 we report the 99% confidence interval of the MC price

obtained with the maximum drift baseline as two red-dashed lines. As we can see, the optimized loss function is compatible with the baseline price. Following the theoretical result of the HJB equation, we can assert that the agent has learnt a sub-optimal policy. Since the LV model differs from the BS one for a corrective term in the HJB equation (49), we expect that the optimal solution in the LV framework will be in a close region of the maximum drift strategy. Thus we train another agent with the direct policy learning by parameterizing its action with the baseline strategy and choosing a smaller learning rate of  $10^{-4}$ . With this parameterization, the RL agent actions, constituting the risky allocation strategy, are computed by summing at each observational time  $T_k \in \mathcal{T}$  the FFNN output with the Equation (36). Even in this case, the corresponding learning curve (solid black line in Figure 4) is stuck in the maximum drift strategy. The first possible reason for this behaviour is that the learning strategy of the direct policy approach is too simple for the learning task. The second interpretation is that the volatility versor in the HJB does not affect significantly the optimum location since it is a second-order term and is lost in the MC error.

Because of that, we change the learning strategy by adopting a more sophisticated one: the PPO algorithm. The big advantage of PPO is that, in addition to the policy, a guess of the value function in Equation (49) is also computed by a parameterization through an FFNN. As for the direct policy approach, we use the BS environment to fine-tune the PPO hyper-parameters. Again we experienced that deeper FFNNs outperform shallow ones and tanh is the most preferable. We report for completeness the values of the other PPO hyper-parameters, for the description of which we refer to [25]: learning rate  $3 \times 10^{-4}$ ,  $\lambda = 0.95$ ,  $\epsilon = 0.2$ ,  $c_1 = 0.7$ ,  $c_2 = 0$  and mini-batch<sup>8</sup> size of 2048 episodes.

We have trained a 5 layer FFNN with 8 neurons each with the PPO algorithm in two different environments: one environment generates the reward according to Equation (55), while the other exploits the reward function (57) where we set  $\gamma = 0.98$  as discount factor to make the agent prefer immediate rewards more than in the previous simulations. Again, for both environments, we train two different agents: one whose action is given by Equation (53), while the other implements the action parameterization with the baseline. We report in Figure 5 the results of the learning curves. In all the PPO cases, the trained agents seem to be stuck in the sub-optimal maximum drift strategy policy. For the case of the immediate reward function (Figure 5 on the right), we can not compare the saturation values reached by the learning curves with the baseline price, since the former are discounted due to  $\lambda, \gamma \neq 1$ . Thus, to compare the performance of PPO with the baseline price in the test phase and ensure consistency in unit of measurement, it is necessary to conduct a MC simulation with frozen agents and set  $\lambda$  and  $\gamma$  values to 1. We report the results in Table 2. The RL agents we have trained with PPO give all prices compatible with the baseline one.

<sup>8</sup> A minibatch is a subset of a larger dataset, which is used to train a machine learning model. The model updates its parameters based on the error calculated from the predictions it makes on each minibatch. This allows for more efficient memory use, faster convergence of the model, and better generalization to new data.



**Fig. 5.** Learning curves of the PPO algorithm applied to the local volatility scenario. On the horizontal axis the number of training episodes. The solid lines are the moving average of the realized rewards on the last  $10^5$  episodes. The shadows represent the 98% confidence intervals. The label *reward (a)* indicates that the RL agent is trained in an environment where the reward function is defined by Equation (55), while *reward (b)* refers to the reward function in Equation (57). The solid black lines are the learning curves of an agent with free actions, while the grey solid ones are the learning curves of an agent whose actions are parameterized with the maximum drift strategy. In both plots, the agents are parameterized by an FFNN with 5 hidden layers with 5 neurons each with tanh activation function. The horizontal grey-dashed lines delimit the 99% confidence interval of the MC price obtained by maximizing path-wise the TVS drift through Equation (36)

## 6 Conclusion and further developments

In this paper, we described a non-trivial control problem related to derivative contracts on target volatility strategies. In particular, we have considered an

**Table 2.** Comparison of TVO prices under local volatility scenario. The baseline price is obtained by applying path-wise the maximum drift strategy (36). The RL free agent implements the policy defined by Equation (53), while the parameterized agent chooses its actions in terms of the baseline strategy. We label with letter (a) the agents trained in environments with the reward function (55), while with (b) the reward function (57). The option parameters are:  $I_0 = K = 1$  [EUR],  $T = 2$  [yr] and  $\bar{\sigma} = 5\%$

Pricing method	TVO price [EUR]
Baseline	$(4.138 \pm 0.005) \times 10^{-2}$
RL free (a)	$(4.127 \pm 0.005) \times 10^{-2}$
RL parameterized (a)	$(4.131 \pm 0.005) \times 10^{-2}$
RL free (b)	$(4.130 \pm 0.005) \times 10^{-2}$
RL parameterized (b)	$(4.135 \pm 0.005) \times 10^{-2}$

option writer selling a call option to a fund manager as protection on the capital invested on the TVS. We showed how the presence of different funding costs coming from hedging the risky assets underlying the TVS, obliges the writer to solve a stochastic optimal control problem to price the protection. This is due to the fact that the option-seller’s strategy is not self-financing. This kind of control problem is hard to solve because here the control process affects both drift and diffusive coefficients of the controlled process. Despite its complexity, our first contribution is the derivation of a closed form solution of the control problem in a Black and Scholes framework, which could represent a useful tool for practitioners since it outperforms intuitive trading strategies. We have derived this solution in two different ways: first by applying the Gyöngy Lemma and then by writing the HJB equation. We numerically studied the problem in the more general local volatility model where the solution is not available and thus a numerical investigation is needed. We tackled the problem by means of the novel RL techniques, by both the direct policy learning and the proximal policy optimization one. We used the BS model, where the solution is *a priori* known, as benchmark to perform a series of fine-tuning of the RL algorithm hyper-parameters, such as the artificial neural network architecture. We have tested in the LV model the two RL approaches and from our simulations, we have evidence that nor the simple direct policy learning strategy nor the sophisticated PPO are able to outperform our analytical solution applied path-wise. Thus our analytical result for the Black and Scholes model seems to be a good proxy solution also for the local volatility one.

This result seems to be a local optimum from the HJB equation of the problem, since in the LV model the volatility versor term should influence the RL agent actions. Thus natural development of this work could be to solve the HJB numerically in low dimension in order to check why such sophisticated algorithms are not able to find the global optimum of the problem, or to understand which are the key elements of the problem, such as market data or the payoff function, that can give rise to a solution far from the intuitive one.

## References

1. Alberverio, S., Victoria, S., and Wallbaum K.: The volatility target effect in investment-linked products with embedded American-type derivatives. *Investment Management and Financial Innovations* **16-3** (2019) 18-28
2. Black, F. and Scholes, M.: The Pricing of Options and Corporate Liabilities. *The Journal of Political Economy* **81-3** (1973) 637-654
3. Chew, L.: Target volatility asset allocation strategy. Society of Actuaries, International News (2011)
4. Deng, Y., Bao, F., Kong, Y., Ren, Z., and Dai, Q.: Deep direct reinforcement learning for financial signal representation and trading. *IEEE Transactions on Neural Networks and Learning Systems* **28-3** (2017) 653-664
5. Derman, E. and Kani, I.: Riding on a Smile. *Risk* **7** (1994) 32-39
6. Di Graziano, G., and Torricelli, L.: Target volatility option pricing. *International Journal of Theoretical and Applied Finance* **15-1** (2012) 1-17
7. Di Persio, L., Prezioso, L., and Wallbaum, K.: Closed-end formula for options linked to target volatility strategies. Arxiv preprint (2019)
8. Dozat, T.: Incorporating Nesterov momentum into Adam. International Conference on Learning Representations (2016)
9. Dupire, B.: Pricing with a smile. *Risk magazine* **7-1** (1994) 18-20
10. Gabrielli, S., Pallavicini, A., and Scoleri, S.: Funding adjustments in equity linear products. *Risk* (2020)
11. Gatheral, J.: The volatility surface: a practitioner's guide. Wiley Finance (2006)
12. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT Press (2016)
13. Grasselli, M., and Romo, J.M.: Stochastic skew and target volatility options. *Journal of Futures Markets* **36-2** (2016) 174-193
14. Gyöngy, I.: Mimicking the one-dimensional marginal distributions of processes having an ito differential. *Probability Theory and Related Fields* **4-71** (1986) 501-516
15. Halperin, I.: Q-Learner in the Black-Scholes(-Merton) worlds. *The Journal of Derivatives* **28-1** (2020) 99-122
16. Hinton, G., Srivastava, N., and Swersky, K.: Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on **8-14** (2012)
17. Hocquard, A., Ng, S., and Papageorgiou, N.: A constant volatility framework for managing tail risk. *The Journal of Portfolio Management* **2-39** (2013) 28-40
18. Kim, Y., and Enke, D.: A dynamic target volatility strategy for asset allocation using artificial neural networks. *The Engineering Economist* **4-63** (2018) 273-290
19. Kolm, P.N., and Ritter, G.: Dynamic replication and hedging: a reinforcement learning approach. *The Journal of Financial Data Science* **1-1** (2019) 159-171
20. Longstaff, F., and Schwartz, E.: Valuing American options by simulation: a simple least-squares approach. *Review of Financial Studies* **14-1** (2001) 113-147
21. Morrison, S., and Tadrowski, L.: Guarantees and target volatility funds. Moody's Analytics (2013)
22. Perchet, R., De Carvalho, R.L., Heckel, T., and Moulin, P.: Predicting the success of volatility targeting strategies: application to equities and other asset classes. *The Journal of Alternative Investments* **3-18** (2016) 21-38
23. Puheim, M., and Madarász, L.: Normalization of inputs and outputs of neural network based robotic arm controller in role of inverse kinematic model. 2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (2014) 85-89

24. Schulman, J., Moritz, P., Levine, S., Jordan, M.I., and Abbeel, P.: High-dimensional continuous control using generalized advantage estimation. 4th International Conference on Learning Representations (2016)
25. Schulman, J., Wolski, P., Dhariwal, P., Radford, A., and Klimov, O.: Proximal policy optimization. Arxiv preprint (2017)
26. Sola, J., and Sevilla, J.: Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on Nuclear Science **44-3** (1997) 1464-1468
27. Sutton, R.S., and Barto, A.G.: Reinforcement learning: an introduction. MIT Press, Cambridge (2018)
28. Xue, Y.: Target Volatility fund: an effective risk management tool for VA?. Society of Actuaries, International News (2012)

**Mathematical Methods in Economics and Finance – m<sup>2</sup>ef**

Vol. 17/18, No. 1, 2022/2023

ISSN print edition: 1971-6419 – ISSN online edition: 1971-3878

Web page: <http://www.unive.it/m2ef/> – E-mail: [m2ef@unive.it](mailto:m2ef@unive.it)