

# Reinforcement Learning for Options on Target Volatility Funds

Roberto Daluise\*    Emanuele Nastasi†    Andrea Pallavicini‡  
Stefano Polo §

First Version: March 2021. This Version: May 18, 2021

## Abstract

In recent years Reinforcement Learning (RL) techniques are gaining popularity in the field of quantitative finance since they are proving to be an efficient way to solve high-dimensional optimal control problems. Our research project is devoted to apply RL to price derivative contracts on target volatility strategies (TVSs), portfolios of risky assets and a risk-free one dynamically rebalanced in order to keep the realized volatility of the portfolio on a certain level. The uncertainty in the TVS risky portfolio composition along with the difference in hedging costs for each component requires to solve a stochastic control problem to evaluate the option prices. The topic of hedging costs is a novelty never dealt in the TVS literature and we provide a formal description of the entire control problem. We tackle the problem by implementing a Reinforcement Learning algorithm to determine the optimal risky portfolio composition leading to the most conservative option price. We investigate the problem for two models of the risky asset dynamics: time-dependent Black & Scholes and Local Volatility. In the first case we prove the existence of an analytical solution of the problem; a result that we use as benchmark to perform a series of fine-tuning of the hyper-parameters of the RL algorithm. At the end we provide numerical results for the Local Volatility model, for which an *a priori* solution is not available.

**JEL classification codes:** C63, C45, C61, G11, G13.

**AMS classification codes:** 65C05, 91G20, 91G60.

**Keywords:** Reinforcement Learning, Hedging Costs, Funding Costs, Proximal Policy Optimization, Target Volatility, Asset Allocation.

---

\*Intesa Sanpaolo Milan, roberto.daluise@intesasanpaolo.com

†Marketz S.p.A., emanuele.nastasi@marketz.eu

‡Imperial College London and Intesa Sanpaolo Milan, a.pallavicini@imperial.ac.uk

§Intesa Sanpaolo Milan, stefano.polo@intesasanpaolo.com

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Target Volatility Strategy</b>	<b>3</b>
2.1	The strategy Price Process . . . . .	4
2.2	The Volatility Targeting Constraint . . . . .	6
<b>3</b>	<b>Derivative Pricing</b>	<b>7</b>
3.1	European Options . . . . .	7
3.2	Stochastic Optimal Control Problem . . . . .	8
3.3	Black and Scholes Model . . . . .	9
3.3.1	Black's Closed form solution . . . . .	10
3.3.2	Active Asset (or Bang Bang) Solution . . . . .	10
3.4	Hamilton-Jacobi-Bellman Equation for Target Volatility Options . . . .	12
<b>4</b>	<b>Reinforcement Learning</b>	<b>13</b>
4.1	Episode Description: Agent-Environment Interaction . . . . .	14
4.2	Policy and Action Parameterization . . . . .	15

# 1 Introduction

In the recent years portfolio managers are exposed to very low interest rates and quickly changing market volatilities. An effective solution to control risks under such 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 portfolio of risky assets (typically equities) and a risk-free asset dynamically re-balanced with the aim of maintaining the overall portfolio volatility level closed to some target value. This products were initially offered in the Asian markets, see for instance the reports of Chew (2011) and Xue (2012) 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 Morrison (2013).

In literature TVSs are tested to investigate their performances in term of realized returns, see for instance Hocquard et al. (2013) and Perchet et al. (2016), and the soundness of the volatility targeting algorithm, as described in Kim and Enke (2018). Moreover, in pricing literature derivative contracts on TVS are studied, see in particular Di Graziano and Torricelli (2012), Grasselli and Romo (2016), and Albeverio et al. (2019).

In this contribution we study a problem related to TVS. We deal with the funding costs coming from hedging the risky assets underlying the TVS. We consider the point of view of a bank selling a protection to a portfolio manager on the capital invested in a TVS. the portfolio manager has the freedom of changing the relative weights of the risky assets during the life of the TVS. Since the risky assets have different hedging costs, the bank shall adjust the price of the protection to include them in the worst-case scenario. Hence, the pricing problem becomes a dynamical control problem over the risky portfolio composition. In our contribution we describe the dynamical control problem, and we derive an analytical solution in the simple case of the risky assets driven by a Black-Scholes (BS) model. Then, we tackle the problem in the general case by using reinforcement learning (RL) algorithms.

The paper is organized as follows. In Section 2 we describe the dynamics of a TVS in presence of valuation adjustments. Then, in Section 3 we derive the analytical results, and in Section 4 we illustrate how we have applied RL to solve the dynamic control problem. We conclude the paper with ?? where we present the numerical results obtained in this work.

## 2 Target Volatility Strategy

In a TVS the fund manager selects an allocation strategy aiming at stabilizing the portfolio volatility to a target level. Clients investing in the fund pay a running fee for the service and their capital is protected. The fund manager usually buys 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 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-bank. The fund manager is not implementing the strategy by trading in the market, and he is not subject to additional costs to access the market. On the other way, the bank is paying such costs since she is actively hedging the call option sold to the manager.

The bank trading activity implemented to actively hedge the option requires funding the collateral procedures of the hedging instruments along with any lending/borrowing fee. The price of a financial product sold by the bank is modified to include any valuation adjustment due to the trading activity. We proceed by defining the price process for 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 representing the actual distribution of supply and demand shocks on equities prices.

We consider a fund trading a basket of  $n$  risky securities with price process  $S_t^i$  with  $i = 1, \dots, n$  funded with a cash account  $B_t$  accruing at  $r_t$ . Any dividend paid by the securities is re-invested in the fund, so that we limit our analysis to total return securities, namely we assume that holding the security is self-financing. Here, we assume that the TVS is implemented in continuous time, even if in the practice we can implement the strategy only on a discrete set of dates. We introduce the deflated gain process  $\bar{G}_t^i$  associated to 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<sup>2</sup> and cumulative dividend processes as

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

<sup>1</sup>Here we neglect discounting factors.

<sup>2</sup>We use bar notation for deflated quantities: processes expressed in terms of  $B_t$ .

where  $\pi_t^i$  represents the cumulative contractual-coupon process paid by the security, and  $\psi_t^i$  represents the cumulative valuation adjustments.

Valuation adjustments (XVAs) is a topic widely discussed in the literature. We refer to Brigo et al. (2013) for a discussion. Since fund managers allocating TVS usually rely on Equity assets, here we use the results of Gabrielli et al. (2020) which 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 given by

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

where  $\phi_t$  are the running fees earned by the fund manager for his 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>3</sup>.

Now, in order to prevent arbitrages, we assume the existence of a risk-neutral measure  $\mathbb{Q}$  under which the deflated gain processes of all traded security 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.

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

where  $M_t^i$  are martingale 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 a cash account rate rate  $r_t$  compensated for the fund manager fees

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

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

---

<sup>3</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  $b \cdot A = \sum_i b_i a_{ij}$ .

adjustment  $\psi_t^i$ , which is specific of the investor pricing the strategy. Thus, the TVS defined in the contract will be

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

leading to the following price process dynamics

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

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 (Equation (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 cost of carry  $\mu_t^i$ .

## 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  $B_t$ . Usually TVSs are total-return products; thus we are justified in assuming  $\pi_t = 0$ . 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 he can trade and after that his choices will be scaled by the automatic target volatility algorithm<sup>4</sup>  $\omega_t$ .

To derive the expression for  $\omega_t$  we need to assume a generic continuous semi-martingales dynamics under the risk-neutral measure for the underlying securities, so that we can write Equation (8) as

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

---

<sup>4</sup>We recall that the universe of assets where the manager can trade and the value of  $\bar{\sigma}$  are written in the contract.

where  $\nu_t$  is an adapted matrix process ensuring the existence of a solution for the SDE 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>5</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 adjustments due to valuation adjustments and fees.

### 3 Derivative Pricing

A derivative contract on the TVS with maturity  $T$  can be defined as

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

where  $D(0, T; \zeta)$  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 strategies since we do not have any information on the future activity of the fund manager.

#### 3.1 European Options

If the derivative contract depends only on the marginal distribution of  $I_t$  at maturity (a European payoff), we are able to prove that exists an optimal strategy, and we are able to calculate it. 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

$$\frac{dI_t^{\text{MP}}}{I_t^{\text{MP}}} := \left( r_t - \ell \left( t, I_t^{\text{MP}} \right) \right) dt + \bar{\sigma} dW_t^{\text{MP}}, \quad (20)$$

---

<sup>5</sup>In all formulae the norm for a vector  $a$  is defined as  $\|a\| := \sqrt{a \cdot a}$ .

where the local drift is defined as

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

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

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

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

**Proposition 3.1.** *A European payoff on the TVS 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  $\mu_t = 0$ .*

**Remark 3.1** (Existence of the Solution in the General Case). *In a more general settings we are not able to find an explicit solution. A proof of the existence of the solution in a general setting is missing. This is a stochastic optimal control problem where by homogeneity we can suppose that  $\alpha_t$  lives in a compact domain, namely (a subset of) the unit simplex. A least if  $r_t$ ,  $\mu_t^i$  and  $\nu_t^i$  are (uniformly) bounded, and if the eigenvalues of  $\nu_t$  are (uniformly) bounded away from zero, then drift and diffusion should be uniformly Lipschitz in  $\alpha_t$ , and classical theorems should exist.*

### 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

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

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  $(X_t, I_t, \alpha_t)$ . We calculate



$I_{T_k}(\alpha_{T_{k-1}})$  for any given strategy  $\alpha_{T_{k-1}}$  by a suitable discretization of (17) starting from  $X_{T_{k-1}}$  and  $I_{T_{k-1}}$ .

Thus the derivative price is given by:

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

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 Model

In the Black and Scholes (1973) 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 can be written as

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

so that the optimization problem can be solved 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 function of time.

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

**Proposition 3.2.** *When the underlying securities follow a Black and Scholes (BS) model with deterministic rates, the optimal strategy for a monotone payoff consists in minimizing the local drift function, independently of the current state  $I_t$*

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

The absence of stochastic elements in Equation (28) makes the optimal strategy known *a priori*; in fact one can solve the optimization problem once for all for each  $t \in \mathcal{T}$  just looking at the market data  $\mu(t)$  and  $\nu(t)$  for the securities. Once  $\alpha^*$  is known, then one can price the payoff by the following BS formula

$$V_0^{BS} = BS(F^{\text{TVS}}(0, T; \alpha^*), K, T, \bar{\sigma}, D(0, T; \zeta)), \quad (29)$$

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

$$F^{\text{TVS}}(t, T; \alpha) = I_t \exp \left[ \int_t^T \left( r(u) - \bar{\sigma} \frac{\alpha(u) \cdot \mu(u)}{\|\alpha(u) \cdot \nu(u)\|} \right) du \right], \quad (30)$$

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

### 3.3.1 Black's 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 3.3.** *Let be  $\mu, \alpha \in \mathbb{R}^n$ ,  $\nu \in \mathbb{R}^{n \times n}$  be a full rank matrix and  $\Sigma := \nu\nu^T$ . Then the closed solution of the optimization problem (28) is*

$$\alpha^* = -\frac{\Sigma^{-1} \cdot \mu}{\|(\Sigma^{-1} \cdot \mu) \cdot \nu\|} \quad (31)$$

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

$$\begin{aligned} & \text{minimize } \alpha \cdot \mu \\ & \text{subject to } \|\alpha \cdot \nu\|^2 = 1 \end{aligned} \quad (32)$$

By setting the Lagrangian function associated with the problem

$$\mathcal{L}(\alpha, \lambda) = \alpha \cdot \mu - \lambda (\|\alpha \cdot \nu\|^2 - 1), \quad (33)$$

we obtain the first order conditions

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \alpha} = \mu - 2\lambda \Sigma \cdot \alpha = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = \|\alpha \cdot \nu\|^2 - 1 = 0, \end{cases} \quad (34)$$

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

$$\alpha^* = \pm \frac{\Sigma^{-1} \cdot \mu}{\|(\Sigma^{-1} \cdot \mu) \cdot \nu\|} \quad (35)$$

We take the minus sign to get the minimum value of the TVS local drift.  $\square$

### 3.3.2 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 are nonnegative and the only constraint on portfolio weights is nonnegativity, which would mean a long only strategy by the fund manager.

**Lemma 3.4.** *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  $\Sigma = \nu\nu^T$ . Then*

$$\inf_{\alpha \in \mathbb{R}_+^n \setminus \{0\}} \frac{\alpha \cdot \mu}{\|\alpha \cdot \nu\|} = \min_{i \leq n} \frac{\mu_i}{\sqrt{\Sigma_{ii}}}; \quad (36)$$

*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 (36) reduces to the maximization of its denominator: the required infimum will be square root of the reciprocal of

$$\sup \left\{ \|\alpha \cdot \nu\|^2 \mid \alpha \in \mathbb{R}_+^n, \alpha \cdot \mathbb{1} = 1 \right\}.$$

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

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

because  $\sum_i \alpha_i = 1$ . Since we trivially have equality for  $\alpha_i = \delta_{i\bar{n}}$ , 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  $\beta = M\alpha$ :

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

which by the first part of the proof equals

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

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

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

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

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

□

### 3.4 Hamilton-Jacobi-Bellman Equation for Target Volatility Options

In this section we want to provide to the reader a formal description of the dynamic problem associated to options on target volatility strategies by writing the Hamilton-Jacobi-Bellman equation for the derivative value. We prove that from this description it is available the same closed formula (28) for the time-dependent BS model derived from the Gyöngy.

In full generality we suppose that the risky securities follows a generic multidimensional diffusive process  $X_t$  given by the following SDE

$$dX_t = M(X_t)dt + \Sigma(X_t) \cdot dW_t. \quad (37)$$

The TVS dynamics is given as follows

$$\frac{dI_t}{I_t} = \left( r_t(X_t) - \frac{\bar{\sigma}\alpha_t}{\|\alpha_t \cdot \nu_t(X_t)\|} \cdot \mu_t(X_t) \right) dt + \frac{\bar{\sigma}\alpha_t}{\|\alpha_t \cdot \nu_t(X_t)\|} \cdot \nu_t(X_t) \cdot dW_t \quad (38)$$

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

$$\begin{aligned} \frac{\partial V}{\partial t} + \max_{\alpha} \left\{ \left( r_t(X) - \bar{\sigma} \frac{\alpha \cdot \mu_t(X)}{\|\alpha \cdot \nu_t(X)\|} \right) I \frac{\partial V}{\partial I} + (\nabla_X V) \cdot M(X) + \frac{1}{2} \bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} \right. \\ \left. + \frac{1}{2} \text{Tr}(\Sigma(X)^T (H_X V) \Sigma(X)) + (\nabla_{X,I} V) \cdot \Sigma(X) \cdot \left( I \bar{\sigma} \frac{\alpha \cdot \nu_t(X)}{\|\alpha \cdot \nu_t(X)\|} \right) \right\} = 0, \end{aligned} \quad (39)$$

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:

$$\nabla_{X,I} V = \left( \frac{\partial^2 V}{\partial X^1 \partial I}, \dots, \frac{\partial^2 V}{\partial X^n \partial I} \right)^T. \quad (40)$$

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

$$\begin{aligned} \frac{\partial V}{\partial t} + r_t(X) I \frac{\partial V}{\partial I} + (\nabla_X V) \cdot M(X) + \frac{1}{2} \bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} + \frac{1}{2} \text{Tr}(\Sigma(X)^T (H_X V) \Sigma(X)) \\ + \max_{\alpha} \left\{ -\frac{\partial V}{\partial I} I \bar{\sigma} \frac{\alpha \cdot \mu_t(X)}{\|\alpha \cdot \nu_t(X)\|} + (\nabla_{X,I} V) \cdot \Sigma(X) \cdot \left( I \bar{\sigma} \frac{\alpha \cdot \nu_t(X)}{\|\alpha \cdot \nu_t(X)\|} \right) \right\} = 0. \end{aligned} \quad (41)$$

Equation (41) represent the Hamilton-Jacobi-Bellman equation describing the TVS dynamic problem for a generic dynamics of the risky securities underlying the portfolio. In our work we have that<sup>6</sup>  $X_t = S_t$ ,  $M(X_t) := (r_t \mathbf{1} - \mu_t) \circ S_t$  and  $\Sigma(X_t) = S_t \circ \nu_t$ .

<sup>6</sup>We use  $\circ$  for the element-wise product between two vectors and between vector and matrix.

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 all zero. Thus the reduced HJB is

$$\frac{\partial V}{\partial t} + Ir(t) \frac{\partial V}{\partial I} + \frac{1}{2} \bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} + \max_{\alpha} \left\{ -\frac{\partial V}{\partial I} I \bar{\sigma} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|} \right\} = 0, \quad (42)$$

so that if the payoff is increasing in  $I_T$  by homogeneity of the SDE we get that  $V$  is increasing and so the solution is given by

$$\alpha^*(t) = \arg \min_{\alpha} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|} \quad (43)$$

which is the same result expressed in eq. (28). If the payoff is decreasing in  $I_T$  then the solution will be the arg max.

Conversely, if we deal with a LV model for the  $S_t$ -dynamics, then  $\nu_t = \nu(t, S_t)$  and there are no apparent simplifications in eq. (41) thus we are not able to derive a closed form solution as in BS. In fact in the LV dynamics we have in the maximum operator the first order term derived from the drift of  $I_t$  (as in BS) and a second order term which takes into account the volatility's smile functions deriving from the  $S_t$ -dynamics.

## 4 Reinforcement Learning

As we have discussed previously in the case of general payoffs or risky securities dynamics, one must resort to numerical approaches to solve the stochastic control problem related to the TVS. The standard approach could be to use classical techniques based on backwards recursion (24)-(25). However their performances may degrade as the dimension  $n$  of the problem increases. In our contribution we adopt a novel technique that is gaining popularity in many scientific branches for solving stochastic optimal control problems: Reinforcement Learning.

Reinforcement Learning is a branch of Machine Learning (ML) which allows an artificial agent to interact with an environment through actions and observations in order to maximize total rewards to achieve specific goals. In RL the agent is not told which actions to take but instead it must discover by trial and error which are the behaviors yielding to the greatest reward by trying them several times. 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 from pre-collected data as opposed to other ML techniques. 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 Deng et al. (2017) as the pioneers in studying self-taught reinforcement trading problems, while to Ritter and Kolm (2018) and Halperin (2017) for hedging derivatives with RL under market frictions.

In our work we adopt as learning strategy the Proximal Policy Optimization (PPO) developed in Schulman, Wolski, et al. (2017) and Schulman, Moritz, et al. (2015). PPO is an actor-critic algorithm well-suited for continuous control problems. The term “actor-critic” means that it estimates the value function  $J$ , which represents the expected value of future rewards, and updates the policy in the direction suggested by  $J$ . In PPO the policy distribution and the value function are parameterized by a set of parameters  $\theta$ :  $J^\theta$ ,  $\pi^\theta$ . The PPO collects a small batch of experiences interacting with the environment to update  $\theta$  by computing an estimator of the policy gradient and plugging it into a stochastic gradient ascent algorithm in order to maximize an objective function  $L^{\text{PPO}}$ . The PPO objective is defined as

$$L^{\text{PPO}} = L^\pi - c_1 L^J + c_2 L^E \quad (44)$$

where  $L^\pi$  is the policy surrogate,  $L^J$  the value function error term and  $L^E$  the entropy bonus, while  $c_1$  and  $c_2$  are hyper-parameters of the algorithm. We refer to Schulman, Wolski, et al. (2017) for a deeper description of those terms. We use the PPO implementation found in OPENAI BASELINES (Dhariwal et al., 2017), where the parameterization procedure is performed by using artificial neural networks (ANNs), where the set of parameters  $\theta$  are the hidden weights of the networks. In the following section we describe the way we have formalized the TVS problem in the Reinforcement Learning framework.

#### 4.1 Episode Description: Agent-Environment Interaction

We consider an episode  $\tau$  of length  $m + 1$  that takes place on a discrete time-grid of fixing times  $\mathcal{T} := \{T_0, \dots, T_k, \dots, T_m\}$  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  $\pi_k^\theta$ . In our case the agent can choose the composition of the risky asset portfolio, so that the policy is the asset allocation weights  $\alpha$ . From the HJB equation we have derived in section Section 3.4, we select as state of the environment the following block

$$s_k := [T_k, I_{T_k}, S_{T_k}] \quad \forall T_k \in \mathcal{T}. \quad (45)$$

This state contains all the informations needed by the agent to take an optimal action, leading to the maximum option price on the TVS. Once the agent has selected the action  $a_k$ , it receives the next time  $T_{k+1}$  a reward  $r_{k+1}$  generated by the environment defined by

$$r_{k+1} := \gamma^k [V_{\text{BS}}(T_{k+1}) - V_{\text{BS}}(T_k)], \quad (46)$$

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

$$V_{\text{BS}}(T_k) := BS(F^{\text{TVS}}(T_k, T; \alpha_{\text{BS}}^*), K, T - T_k, \bar{\sigma}, D(T_k, T; \zeta)) \quad \text{and} \quad V(T_0) = 0 \quad (47)$$

with  $\alpha_{BS}^*$  the BS optimal strategy (28) calculated in the state  $s_k$ . The parameter  $\gamma$  plays the role of a discount rate 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. Choosing  $\gamma < 1$  introduces a bias in the final result. If we take the cumulative reward per episode and set  $\gamma = 1$  we obtain

$$R(\tau) = \sum_{k=0}^{m-1} r_{k+1} = \sum_{k=0}^{m-1} [V_{BS}(T_{k+1}) - V_{BS}(T_k)] = V_{BS}(T) = (I_T - K)^+, \quad (48)$$

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

Thus one can train the agent choosing  $\gamma \in [0, 1]$ , and then run, as test phase, a Monte Carlo (MC) simulation with  $\gamma = 1$ , where, if the agent has reached the optimal policy  $\pi^* = \pi^{\theta^*}$ , the average of  $R(\tau)$  along different episodes will match the optimal undiscounted price of the derivative contract on the TVS.

## 4.2 Policy and Action Parameterization

As discussed before, PPO parameterizes the agent policy by means of an artificial neural network. In our case we use a feed forward neural network (FFNN), a network with an acyclic topology and completely characterized by the following hyper-parameters: number of hidden layers  $N$ , number of neurons per hidden layer  $M$ , activation function per hidden layer and output activation function. The number of output neurons is fixed by the action space dimension, which is the number of assets  $n$  in the risky portfolio, while the number of input neurons is determined by the dimensionality of the state (45),  $n + 2$ .

## References

- Albeverio, S., S. Victoria, and K. Wallbaum (July 2019). “The volatility target effect in investment-linked products with embedded American-type derivatives”. In: *Investment Management and Financial Innovations* 16, pp. 18–28.
- Black, Fischer and Myron Scholes (May 1973). “The Pricing of Options and Corporate Liabilities”. In: *Journal of Political Economy* 81 (3). DOI: [10.2307/1831029](https://doi.org/10.2307/1831029). URL: <http://gen.lib.rus.ec/scimag/index.php?s=10.2307/1831029>.
- Brigo, Damiano, Massimo Morini, and Andrea Pallavicini (Apr. 2013). *Counterparty Credit Risk, Collateral and Funding: With Pricing Cases For All Asset Classes*. Wiley, Chichester.
- Chew, Yuhong (Apr. 2011). *Target Volatility Asset Allocation Strategy*. Tech. rep. Society of Actuaries.
- Deng, Y. et al. (2017). “Deep Direct Reinforcement Learning for Financial Signal Representation and Trading”. In: *IEEE Transactions on Neural Networks and Learning Systems* 28.3, pp. 653–664. DOI: [10.1109/TNNLS.2016.2522401](https://doi.org/10.1109/TNNLS.2016.2522401).
- Dhariwal, Prafulla et al. (2017). *OpenAI Baselines*. <https://github.com/openai/baselines>.
- Di Graziano, Giuseppe and Lorenzo Torricelli (2012). “Target Volatility Option Pricing”. In: *International Journal of Theoretical and Applied Finance* 15.01.
- Gabrielli, Stefania, Andrea Pallavicini, and Stefano Scoleri (June 2020). “Funding Adjustments in Equity Linear Products”. In: *Risk*.
- Grasselli, Martino and Jacinto Marabel Romo (2016). “Stochastic Skew and Target Volatility Options”. In: *Journal of Futures Markets* 36.2, pp. 174–193.
- Halperin, Igor (Dec. 2017). “QLBS: Q-Learner in the Black-Scholes(-Merton) Worlds”. In: *SSRN Electronic Journal*. DOI: [10.2139/ssrn.3087076](https://doi.org/10.2139/ssrn.3087076).
- Hocquard, Alexandre, Sunny Ng, and Nicolas Papageorgiou (2013). “A Constant-Volatility Framework for Managing Tail Risk”. In: *The Journal of Portfolio Management* 39.2, pp. 28–40.
- Kim, Youngmin and David Enke (2018). “A dynamic target volatility strategy for asset allocation using artificial neural networks”. In: *The Engineering Economist* 63.4, pp. 273–290.
- Morrison Steven; Tadrowski, Laura (Sept. 2013). *Guarantees and Target Volatility Funds*. Tech. rep. Moody’s Analytics.
- Perchet, Romain et al. (Jan. 2016). “Predicting the Success of Volatility Targeting Strategies: Application to Equities and Other Asset Classes”. In: *The Journal of Alternative Investments* 18, pp. 21–38.
- Ritter, Gordon and Petter Kolm (Jan. 2018). “Dynamic Replication and Hedging: A Reinforcement Learning Approach”. In: *SSRN Electronic Journal*. DOI: [10.2139/ssrn.3281235](https://doi.org/10.2139/ssrn.3281235).



- Schulman, John, Philipp Moritz, et al. (June 2015). “High-Dimensional Continuous Control Using Generalized Advantage Estimation”. In: *arXiv preprint arXiv:1506.02438*.
- Schulman, John, Filip Wolski, et al. (2017). “Proximal Policy Optimization Algorithms”. In: *CoRR* abs/1707.06347. arXiv: 1707.06347. URL: <http://arxiv.org/abs/1707.06347>.
- Xue, Yuhong (Oct. 2012). *Target Volatility Fund: An Effective Risk Management Tool for VA?* Tech. rep. Society of Actuaries.

DRAFT