Reinforcement Learning for Options on Target Volatility Funds

Roberto Daluiso* Emanuele Nastasi† Andrea Pallavicini‡
Stefano Polo§

May 28, 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 RL 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 and 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 RL hyper-parameters. 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.daluiso@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

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 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, known as target volatility options (TVOs), are studied, see in particular Di Graziano and Torricelli (2012), Grasselli and Romo (2016), and Albeverio et al. (2019).

In this contribution we face a new topic never dealt with in the TVS literature: the funding costs coming from hedging the risky assets underlying the TVS. We consider the point of view of a bank selling a call option to a portfolio manager as protection 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, i.e. the most expensive strategy from the financing costs point of view. 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 (Black and Scholes, 1973). Then, we tackle the problem in the general case of a local volatility (LV) model (Dupire, 1994; Derman and Kani, 1994) by using Reinforcement Learning (RL) algorithms, in particular the proximal policy optimization developed in Schulman, Wolski, et al. (2017).

The paper is organized ad follows. In Section 2 we describe the dynamics of a TVS in presence of valuation adjustments. Then, in Section 3 we study the case of derivative contracts on TVSs, we describe the dynamical control problem giving the expression of the Hamilton-Jacobi-Bellman equation and we derive the analytical results for plain vanilla options under BS dynamics. 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.

A part of this work has been developed during the Master thesis of one of us (SP), where we had the opportunity to collaborate with Marco Bianchetti and Diego Pierluigi Giovannini from Intesa Sanpaolo, Milan, that we tanks. Moreover we wish to thank the Italian computing centre Cineca, which approved our ISCRA C project and provided us the high-computing resources of Marconi100 for the numerical simulations of this contribution.

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[1]

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

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

$$A_t = K + (V_t - K)^+. \tag{2}$$

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 highlights the impact of valuation adjustments.

2.1 The strategy Price Process

We work on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geqslant 0}, \mathbb{P})$ 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^i_t 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 $\bar{G}^i_t$ associated to the risky securities as given by

$$\bar{G}^i_t \coloneqq \bar{S}^i_t + \bar{D}^i_t, \tag{3}$$

where we define the deflated price[2] and cumulative dividend processes as

$$\bar{S}^i_t \coloneqq \frac{S^i_t}{B_t}, \quad \bar{D}^i_t \coloneqq \int_0^t \frac{d\pi^i_u}{B_u} + \int_0^t \frac{d\psi^i_u}{B_u}, \tag{4}$$

where π^i_t represents the cumulative contractual-coupon process paid by the security, and ψ^i_t represents the cumulative valuation adjustments.

[1] Here we neglect discounting factors.

[2] We use bar notation for deflated quantities: processes expressed in terms of B_t.

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

$$\psi_t^i := \int_0^t S_u^i \mu_u^i du, \tag{5}$$

where we call μ_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

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

where ϕ_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

$$d\bar{G}_t^V = q_t \cdot d\bar{G}_t, \tag{7}$$

where q_t^i is the quantity invested in the i-th security[3].

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.

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

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

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

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

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

leading to the following price process dynamics

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

[3] 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^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

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

where we can see the dependency on cost of carry μ_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 α_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

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

where $\mathbb{1}$ 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

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

where $\bar{\sigma}$ is the target volatility value. In practice, this means that the fund manager will select a risky-portfolio choosing α_t equities from the universe where he can trade and after that his choices will be scaled by the automatic target volatility algorithm[4] ω_t.

To derive the expression for ω_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

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

where ν_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 ω_t, and we get[5]

$$\omega_t = \frac{\bar{\sigma}}{\|\alpha_t \cdot \nu_t\|}. \tag{16}$$

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

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

where we can see, as expected, that the strategy grows at the risk-free rate but for adjustments due to valuation adjustments and fees.

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

[5] In all formulae the norm for a vector a is defined as $\|a\| := \sqrt{a \cdot a}$.

3 Derivative Pricing

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

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

where $D(0,T;\zeta)$ is the discount factor with rate ζ_t, inclusive of the derivative valuation adjustments, and π_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

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

where Φ 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^{MP}, and we get by applying the Gyöngy Lemma (Gyöngy, 1986)

$$\frac{dI_t^{\mathrm{MP}}}{I_t^{\mathrm{MP}}} := \left(r_t - \ell_\alpha \left(t, I_t^{\mathrm{MP}} \right) \right) dt + \bar{\sigma} dW_t^{\mathrm{MP}}, \tag{20}$$

where the local drift is defined as

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

and W_t^{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^{MP}, namely

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

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 α_t lives in a compact domain, namely (a subset of) the unit simplex. A least if r_t, μ_t^i and ν_t^i are (uniformly) bounded, and if the eigenvalues of ν_t are (uniformly) bounded away from zero, then drift and diffusion should be uniformly Lipschitz in α_t, and classical theorems should exist (Fleming and Soner, 2006).*

3.2 Stochastic Optimal Control Problem

In presence of valuation adjustments we need to solve the full optimization problem. We discretize the optimal strategy α_t as

$$\alpha_t \coloneqq \sum_k \mathbf{1}_{\{t\in[T_{k-1},T_k)\}}\alpha_{T_{k-1}}, \tag{23}$$

according to a time grid $\mathcal{T} := \{T_0, ..., T_k, ..., T_m\}$ 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 α_t at time T_{k-1} as

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

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, α_t). We calculate $I_{T_k}\left(\alpha_{T_{k-1}}\right)$ 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:

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

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

$$V_{T_m} = \Phi\left(I_{T_m}\right). \tag{26}$$

3.3 Black and Scholes Model

In the 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 α_t^* which depends in principle only on I_t. As a consequence, the local drift can be written as

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

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 non-decreasing European payoff consists in minimizing the local drift function, independently of the current state I_t*

$$\alpha^*(t) \coloneqq \arg\min_{\alpha} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|}. \tag{28}$$

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

$$\alpha^*(t) \coloneqq \arg\max_{\alpha} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|}. \tag{29}$$

The absence of stochastic elements in Equation (28) makes the optimal strategy known *a priori* with no simulation needed; in fact one can solve the optimization problem once for all $\forall t \in \mathcal{T}$ just looking at the market data $\mu(t)$ and $\nu(t)$ for the securities. Once α^* is known, then one can price the payoff by the following BS formula

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

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

$$F^{\mathrm{TVS}}(t,T;\alpha^*) = I_t \exp\left[\int_t^T \left(r(u) - \ell_{\alpha*}\left(u\right)\right) du\right] \quad \text{with} \quad \ell_{\alpha*}(t) = \bar{\sigma}\frac{\mu\left(t\right)\cdot\alpha^*\left(t\right)}{\left\|\alpha^*\left(t\right)\cdot\nu\left(t\right)\right\|}, \tag{31}$$

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 σ and discount factor D.

3.3.1 Optimal Strategy Closed Form

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 $\Sigma := \nu\nu^{\intercal}$. Then the closed solution of the optimization problem (28) is*

$$\alpha^* = -\frac{\Sigma^{-1}\cdot\mu}{\left\|\left(\Sigma^{-1}\cdot\mu\right)\cdot\nu\right\|} \tag{32}$$

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

$$\begin{aligned} &\text{minimize } \alpha\cdot\mu \\ &\text{subject to } \left\|\alpha\cdot\nu\right\|^2 = 1 \end{aligned} \tag{33}$$

By setting the Lagrangian function associated with the problem

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

we obtain the first order conditions

$$\begin{cases} \dfrac{\partial\mathcal{L}}{\partial\alpha} = \mu - 2\lambda\Sigma\cdot\alpha = 0 \\ \dfrac{\partial\mathcal{L}}{\partial\lambda} = \left\|\alpha\cdot\nu\right\|^2 - 1 = 0\,, \end{cases} \tag{35}$$

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

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

We take the minus sign to get the minimum value of the TVS local drift. □

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 $\Sigma = \nu\nu^\intercal$. Then*

$$\inf_{\alpha\in\mathbb{R}^n_+\setminus\{0\}} \frac{\alpha\cdot\mu}{\|\alpha\cdot\nu\|} = \min_{i\leqslant n} \frac{\mu_i}{\sqrt{\Sigma_{ii}}}; \tag{37}$$

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

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 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 Σ is positive definite, hence $\Sigma_{ij} < \sqrt{\Sigma_{ii}\Sigma_{jj}} \leqslant \Sigma_{\bar{\imath}\bar{\imath}}$, which implies

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

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

Next, let us consider the case in which all components of μ are strictly positive, and define M as the diagonal matrix with diagonal μ. 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

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

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

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

One can easily note that as $\epsilon\to 0$, f_ϵ 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

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

□

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 one can recover the same closed formula (28) for the time-dependent BS model which was derived above from the Gyöngy Lemma.

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

$$dX_t = M(X_t)dt + \Sigma(X_t) \cdot dW_t. \tag{38}$$

The TVS dynamics is given as follows

$$\frac{dI_t}{I_t} = \left(r_t(X_t) - \frac{\bar{\sigma}\alpha_t}{\|\alpha_t \cdot \nu_t(X_t)\|} \cdot \mu_t(X_t) \right) dt + \frac{\bar{\sigma}\alpha_t}{\|\alpha_t \cdot \nu_t(X_t)\|} \cdot \nu_t(X_t) \cdot dW_t \tag{39}$$

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

$$\begin{aligned} &\frac{\partial V}{\partial t} + \max_{\alpha} \left\{ \left(r_t(X) - \bar{\sigma} \frac{\alpha \cdot \mu_t(X)}{\|\alpha \cdot \nu_t(X)\|} \right) I \frac{\partial V}{\partial I} + (\nabla_X V) \cdot M(X) + \frac{1}{2}\bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} \right. \\ &\left. + \frac{1}{2} \operatorname{Tr}\left(\Sigma(X)^\intercal (H_X V) \Sigma(X)\right) + (\nabla_{X,I} V) \cdot \Sigma(X) \cdot \left(I\bar{\sigma} \frac{\alpha \cdot \nu_t(X)}{\|\alpha \cdot \nu_t(X)\|} \right) \right\} = 0, \end{aligned} \tag{40}$$

where $\operatorname{Tr}(A)$ 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:

$$\nabla_{X,I} V = \left(\frac{\partial^2 V}{\partial X^1 \partial I}, \dots, \frac{\partial^2 V}{\partial X^n \partial I} \right)^\intercal. \tag{41}$$

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

$$\begin{aligned} &\frac{\partial V}{\partial t} + r_t(X) I \frac{\partial V}{\partial I} + (\nabla_X V) \cdot M(X) + \frac{1}{2}\bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} + \frac{1}{2} \operatorname{Tr}\left(\Sigma(X)^\intercal (H_X V) \Sigma(X)\right) \\ &+ \bar{\sigma} I \max_{\alpha} \left\{ - \frac{\partial V}{\partial I} \frac{\alpha \cdot \mu_t(X)}{\|\alpha \cdot \nu_t(X)\|} + (\nabla_{X,I} V) \cdot \Sigma(X) \cdot \left(\frac{\alpha \cdot \nu_t(X)}{\|\alpha \cdot \nu_t(X)\|} \right) \right\} = 0. \end{aligned} \tag{42}$$

Equation (42) 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[6] $X_t = S_t$, $M(X_t) := (r_t \mathbb{1} - \mu_t) \circ S_t$ and $\Sigma(X_t) = S_t \circ \nu_t$.

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

$$\frac{\partial V}{\partial t} + I r(t) \frac{\partial V}{\partial I} + \frac{1}{2}\bar{\sigma}^2 I^2 \frac{\partial^2 V}{\partial I^2} + \bar{\sigma} I \max_{\alpha} \left\{ - \frac{\partial V}{\partial I} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|} \right\} = 0, \tag{43}$$

[6] We use $\circ$ for the element-wise product between two vectors and between vector and matrix.

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

$$\alpha^*(t) = \arg\min_{\alpha} \frac{\alpha \cdot \mu(t)}{\|\alpha \cdot \nu(t)\|} \tag{44}$$

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. (42) 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 π 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 Kolm and Ritter (2019) and Halperin (2020) 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. (2016). 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 θ: J^θ, π^θ. The PPO collects a small batch of experiences interacting with the environment to update θ by computing an estimator of the policy gradient and plugging it into a stochastic gradient ascent (SGA) algorithm in order to maximize an objective function L^{PPO}. The PPO objective is defined as

$$L^{\text{PPO}} = L^\pi - c_1 L^J + c_2 L^E \tag{45}$$

where L^π is a surrogate of the policy's average reward, 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), and the set of parameters θ 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 τ of length $m+1$ that takes place on a discrete time-grid of fixing times expressed in year fractions $\mathcal{T} := \{T_0, ..., T_k, ..., T_m\}$ 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 π_k^θ. In our case the agent can choose the composition of the risky asset portfolio, so that the policy is the asset allocation weights α. From the HJB equation we have derived in section Section 3.4, we select as state of th environment the following block

$$s_k := [T_k, I_{T_k}, S_{T_k}] \quad \forall T_k \in \mathcal{T}. \tag{46}$$

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. We have defined the reward function in two different ways. A first definition is

$$r_{k+1} = \begin{cases} (I_T - K)^+ & \text{if } T_{k+1} = T \\ 0 & \text{otherwise} \end{cases} \tag{47}$$

Therefore, during the 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 whole episode, leading to a slow learning. However, once the agent has learnt the optimal policy $\pi^* = \pi^{\theta^*}$, the average cumulative reward per episodes will coincide to the optimal TVO price. The second definition of reward function for this problem we have developed is

$$r_{k+1} := \gamma^k [V_{\text{BS}}(T_{k+1}) - V_{\text{BS}}(T_k)], \tag{48}$$

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

$$V_{\text{BS}}(T_k) := BS(F^{\text{TVS}}(T_k, T; \alpha^*_{\text{BS}}, K), K, T - T_k, \bar{\sigma}, D(T_k, T; \zeta)) \quad \text{and} \quad V(T_0) = 0 \tag{49}$$

with α^*_{BS} the BS optimal strategy (28) calculated in the state s_k. This second definition of the reward function does not suffer of nil rewards at intermediate times of the episode. The hyper-parameter γ plays the role of a discount rate in the sense that, as γ approaches to zero, the RL agent will tend to maximize immediate rewards while neglecting possible larger rewards in the future. Choosing[7] $\gamma, \lambda < 1$ and introduces a bias in the final result. If we take the cumulative reward per episode and set $\gamma = \lambda = 1$ we obtain

$$R(\tau) = \sum_{k=0}^{m-1} r_{k+1} \underset{\gamma=\lambda=1}{=} \sum_{k=0}^{m-1} [V_{\text{BS}}(T_{k+1}) - V_{\text{BS}}(T_k)] = V_{BS}(T) = (I_T - K)^+, \tag{50}$$

[7] We refer to Schulman, Moritz, et al. (2016) for a more detailed description of λ.

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 $\gamma, \lambda \in [0, 1]$, and then run, as test phase, a Monte Carlo (MC) simulation with $\gamma = \lambda = 1$, 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 (46), $n + 2$. We work with a stochastic policy whose functional form 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 external parameter

$$\pi^\theta(s_k) \sim \mathcal{N}(\mu^\theta(s_k), e^{\log \sigma}). \tag{51}$$

In the OpenAI Baselines implementation of PPO, the parameter $\log \sigma$ is state-independent, but it decreases as the number of 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.

Another important topic is how to parameterize the agent action in order to satisfy some specific constraints on the risky asset allocation weights. In fact, if on the one hand with free α it is sufficient to set bounds in the action space, on the other hand the presence of constraints on the risky strategy imposes the need to parameterize the output of the ANN. In this work we study the constraint where at time $T_k \in \mathcal{T}$ the agent can invest in the risky market adopting long positions only:

$$\alpha_{T_k} \cdot \mathbb{1} = 1 \quad \alpha^i_{T_k} \geqslant 0 \quad \forall i = 1, \dots, n. \tag{52}$$

We study two different action parameterizations for this kind of constraint.

- **$(n-1)$-spherical angles**: we rewrite $\alpha^j_{T_k} = (x^i_k)^2$ and we use $(n-1)$-spherical coordinates with unitary radius. Thus the strategy is parameterized by a vector of angles $\phi_k = (\phi^1_k, \dots, \phi^{n-1}_k)^\intercal$ as follows

$$\begin{aligned}
\alpha^1_{T_k} &:= \cos^2(\phi^1_k), \\
\alpha^2_{T_k} &:= \sin^2(\phi^1_k)\cos^2(\phi^2_k), \\
&\vdots \\
\alpha^{n-1}_{T_k} &:= \sin^2(\phi^1_k)\dots\sin^2(\phi^{n-2}_k)\cos^2(\phi^{n-1}_k), \\
\alpha^n_{T_k} &:= \sin^2(\phi^1_k)\dots\sin^2(\phi^{n-2}_k)\sin^2(\phi^{n-1}_k).
\end{aligned} \tag{53}$$

where $\phi_k^i \in [0, \pi/2]$ to avoid oscillations in α that could slow down the training. Thus, with this parameterization, the ANN sample a vector of angles ($a_k = \phi_k$) and thus the action space is reduced by one dimension.

- **Normalized action**: we restrict the agent actions to a subset of $(\mathbb{R}^+)^n$ and then normalize the action to obtain the corresponding risky allocation strategy through the formula

$$\alpha_{T_k} = \frac{a_k}{a_k \cdot \mathbb{1}}. \tag{54}$$

5 Numerical Investigations

In this section we present the numerical results of our work. In particular we focus our analysis on an European option on the TVS with at-the-money strike considering two different models for the dynamics followed by the risky securities: Black and Scholes and local volatility. In Section 3.3 we have proved that under the BS dynamics the solution of the control problem is given by Equation (28). 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, the availability of this faster solution allows us to identify the key parameters of the ANN, PPO and the action parameterization, and to analyze how they impact on the final results and performances. Then we dedicate a full section to the LV case where the dynamical control problem cannot be solved without the RL agent.

In our work we keep fixed to the default values of OpenAI Baselines some PPO hyperparameters that will not be subject to our fine tuning tests: the SGA learning rate $\rho = 3 \times 10^{-4}$, the lambda Generalized Advantage Estimation parameter $\lambda = 0.95$ and the L^π clip threshold[8] $\epsilon = 0.2$. In every training phase, PPO builds two FFNNs with same architecture and activation functions to evaluate respectively the actions and the value function at each episode step k. The library initializes the hidden weights θ using an orthogonal initialization with $\sqrt{2}$ scaling factor and updates them after collecting a mini-batch of 2048 episodes. The network biases are set to zero.

The two FFNNs take in input the environment state (46). It is well known in literature that ANNs give better performances in the training phase if the input data are well normalized (Sola and Sevilla, 1997; Puheim and Madarász, 2014). Thus we modify the state block by replacing $\forall T_k \in \mathcal{T}$ the TVS spot price I_{T_k} with I_{T_k}/I_0 and the equity price vector S_{T_k} with:

$$X_{T_k} := \frac{\log\left(S_{T_k}/F(0,T_k)\right) + \frac{1}{2}\int_0^{T_k}(\nu^2(t)\cdot\mathbb{1})dt}{\sqrt{\int_0^{T_k}(\nu^2(t)\cdot\mathbb{1})dt}}, \quad X_{T_k} \in [-2.5, 2.5]^n, \tag{55}$$

where $F(t,T)$ is the forward curve vector of the risky assets from t to T. The normalization procedure (55) allows to remap S_{T_K} into a standard Gaussian variable for the BS model, while into a proxy of it in the case of a LV dynamics. We adopt this normalization without loss of generality as the variables involved are observables (forward curve and at-the-money volatilities), since they are values that are always accessible by the fund manager in the market.

[8] We refer to Schulman, Wolski, et al. (2017) for a more accurate description of ϵ and to Schulman, Moritz, et al. (2016) for λ.

All the training experiments required on average 40 CPU cores, 1 NVIDIA Volta V100 GPU and 100 GB RAM to manage the parallel environments generated by OpenAI Baselines; the code is written in Python.

5.1 Black and Scholes: Hyper-Parameters Fine Tuning

We use the BS environment as toy model to understand which parameters of the RL algorithm play key roles in the training and testing phase. We also study which action parameterization is most effective in the case of a long only allocation strategy α. We consider a target volatility call option with the following contract details between bank and fund manager

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

The reward function adopted in this environment is that expressed by Equation (47). We test the action parameterizations discussed in Section 4.2 increasing the number n of equities in the basket; we have evidence that the $(n-1)$-sphere parameterization (53) suffers of labelling problems when $n > 2$. In fact if we consider a spherical case ($n = 3$), eq. (53) introduces an asymmetry in the action space (Figure 1 on the left) that translates into the choice by the agent of sub-optimal policies during the training phase. Because of that we privilege the action parameterization through normalization (54). In Figure 1 on the right we provide the learning curve of a 5×8 ANN[9] with tanh as activation function, $\gamma = 1$, $c_1 = 0.7$ and $c_2 = 0$. This learning curve, just like all the ones we will show in this work, is the best result in term of reward on the last 10^7 episodes among four simulations with different initial seed for the weights θ. This is done since the objective function is not convex.

In Figure 1 we can observe that the average cumulative reward converges as the number of training episode increases to the optimal option value calculated *a priori* with the BS formula (30), proving a successful learning of the optimal BS strategy. This can be observed also by fixing a never-seen MC scenario and interrogating the trained agent on this test path: we see that the ANN takes actions that are close to the analytical allocation strategy (Figure 2).

We use the above PPO parameters since they are the results of our tests of fine tuning (Figure 3). In particular we have trained the agent in BS environments with simple market data structures building a grid search on different hyper-parameters: the ANN architecture, the activation function and the value function trade-off c_1. We first fix a 2×3 ANN and $c_1 = 0.5$ and study the activation function, where we see that the tanh outperforms the popular sigmoid and RELU that reaches suboptimal policies. Then we perform a search in the c_1 direction, where the optimal value is 0.7. At the end we study the number of layers and neurons in the FFNN; we can see in Figure 3 top-right that adding hidden layers in the ANN architecture gives better results in the learning phase than adding hidden neurons per layer, since deeper ANNs seem less sensitive to get stuck in locally optimal strategies. We do not need to turn on the entropy element c_2 since in the BS environment the RL agent finds successfully π^*, and adding an entropy term could deteriorate the learning phase.

All those numerical analysis will be useful to study the LV environment.

[9] We use the notation $L \times N$ to describe a FFNN with L hidden layers each made up of N neurons.

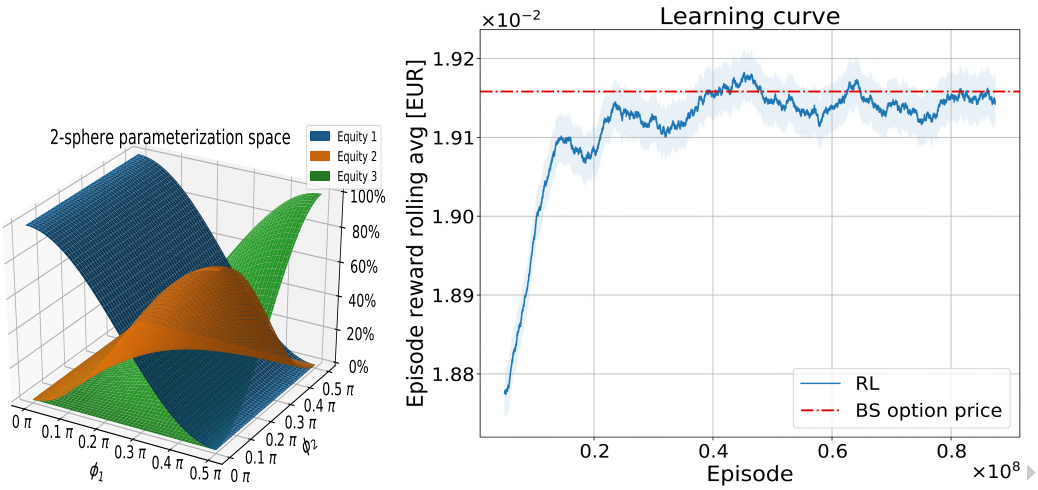

Figure 1: On the left n-sphere strategy parameterization for 3 equities: α in function of the colatitude ϕ_1 and latitude ϕ_2. The domain of the angles is $[0, \pi/2]$. On the right the learning curve for an agent policy parameterized with 5×8 and tanh as activation function. The agent is trained in a BS environment with three equities. On the horizontal axis the number of training episodes. The solid lines are the moving average of the realized rewards on the last 5×10^6 episodes. The shadows represent the 98% confidence intervals. The dash-dotted horizontal red line is the conservative TVO price computed according to the optimal risky portfolio composition through eq. (30).

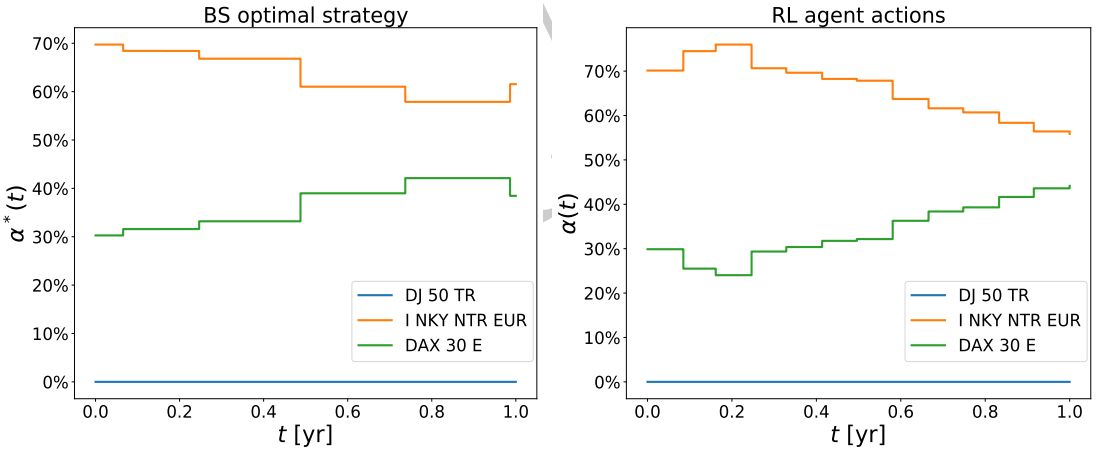

Figure 2: Comparison between the BS optimal risky strategy (Equation (28)) and the RL actions taken in a test episode never seen by the agent. On the horizontal axis time in term of year fractions. The agent training session is showed in Figure 1.

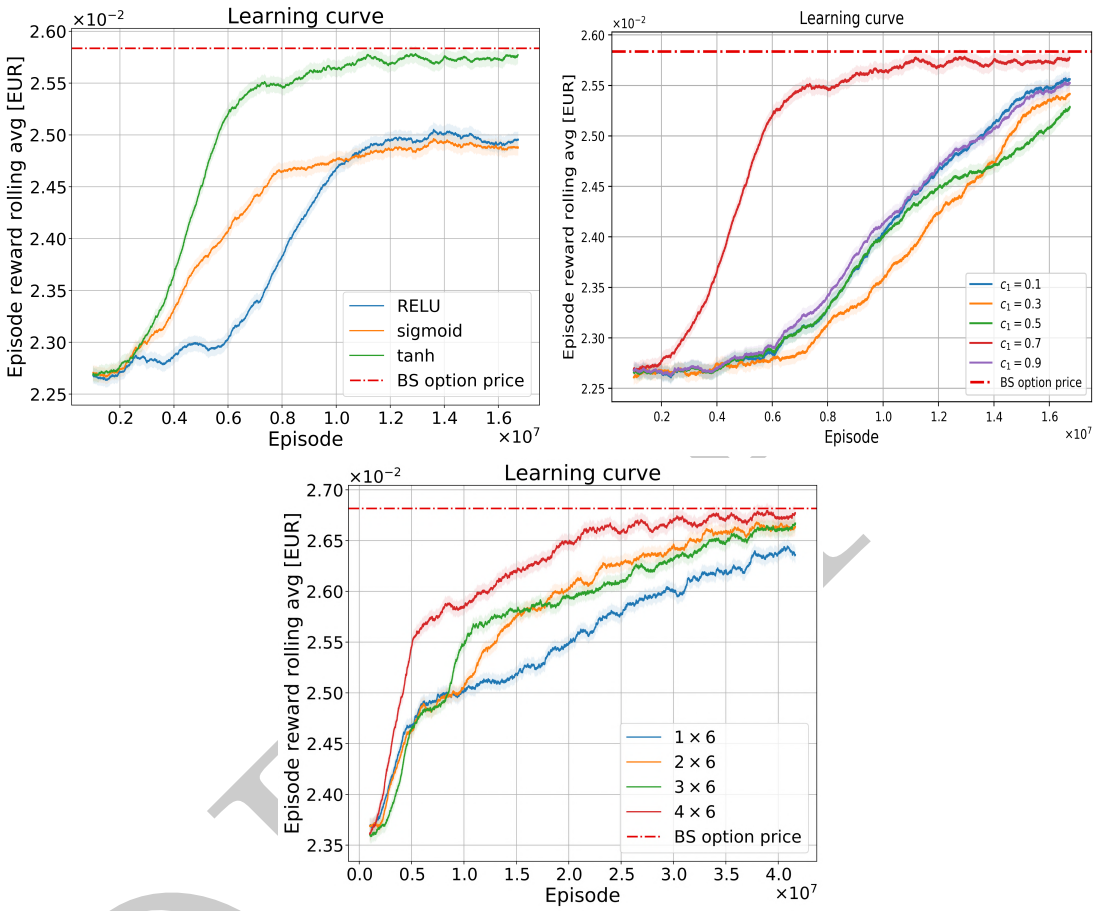

Figure 3: Fine tuning simulations: on the top-left learning curves with different activation functions, on the top-right grid search on the policy-value trade-off and on bottom-centre learning curves on ANNs architectures. On the horizontal axis the number of training episodes. The solid lines are the moving average of the realized rewards on the last 5×10^5 episodes. The shadows represent the 98% confidence intervals. The dash-dotted horizontal red lines are the conservative TVO prices computed according to the optimal risky portfolio composition through eq. (30).

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; in fact if we consider the whole equity smiles then the second order term in the HBJ equation (42) is non null and thus a closed formula for α^* is not available anymore. Because of that one must resort on numerical techniques to recover the problem solution. Unlike BS model where the strategy depends only on time, in LV dynamics there are no unnecessary information provided by the state block in Equation (46) to take the optimal action.

We first apply our RL algorithm in the same market data conditions as in the case showed in Figure 1 on the right, but in the LV dynamics. This choice is done to compare the optimal solution in two well-known financial models. We keep the same ANN architecture as in previous section: a 5×8 FFNN with tanh activation function. This enables us to perform the training using two different initial guesses for the PPO:

- **Random starting point**: the hidden weights θ of the ANN are randomly initialized so that the training phase starts from a casual policy.
- **Black and Scholes initial guess**: the algorithm initialize the values of θ with those of the ANN trained in the BS environment.

The idea behind the second guess is that the LV dynamics differs from BS for a corrective term in the HJB equation and thus we guess that a network which has learnt the BS optimal strategy will be a good starting point for the RL algorithm as sub-optimal value. We present in Figure 4 on the left the resulting learning curves for two ANNs trained with the two initialization methods just described.
Analyzing the training results, we observe that the two curves converges both to the same saturation value which is compatible with the optimal BS price for the TVO curve. The RL agent has learnt in both cases the optimal allocation strategy of the BS model. In fact if we plot the agent actions taken in never-seen scenarios (Figure 4 on the right), we see that they are in accordance with those of BS.

As told before, in the LV model the optimal strategy is not available *a priori* as in BS and thus we can not say if the solution found by the RL agent is sub-optimal. However we can check if the agent is stuck in a local optimum by building a näif strategy to measure the performance of the RL algorithm; we call this strategy "baseline". Since we have proved in Section 3.3 that in the BS model the solution is the allocation strategy that maximizes the local drift of the TVS, we choose as baseline a manager that applies path-wise the BS solution (28). In Table 1 we compare the MC prices obtained applying in 10^6 paths the baseline and the ANNs trained before. All the prices computed with different strategies are 2σ-compatible among them. Thus we can assert that in the market data conditions we have considered, the LV optimal strategy coincides with the BS one. Moreover we observe that the baseline strategy and the BS optimal one provide the same TVO price, such as the risky securities dynamics is irrelevant for the resulting price.

Thus we try different configurations for the hedging costs μ_t and the implied volatility (IV) surfaces in order to find a market where the TVO prices reached by the baseline and by the BS optimal strategy differ between them. From our analyzes we have evidence that in markets

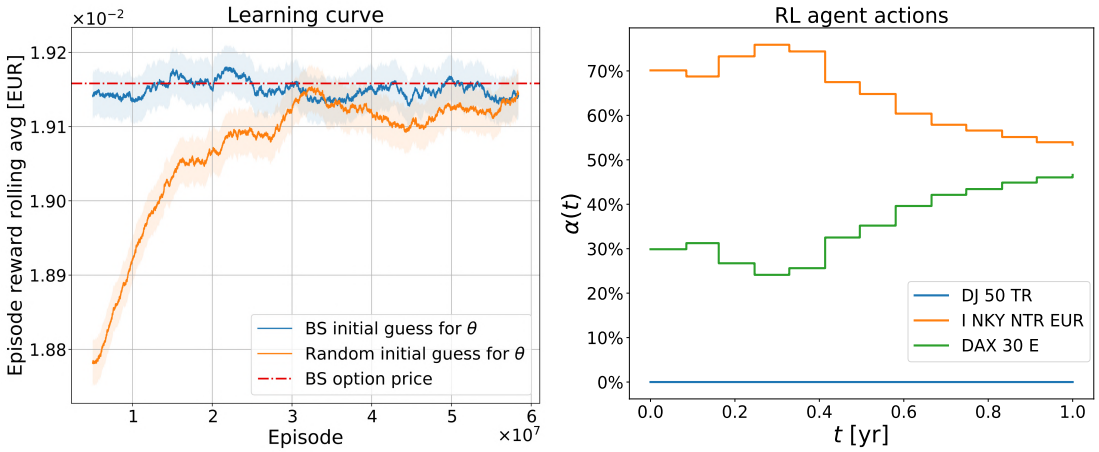

Figure 4: On the left, learning curve for two 5×8 FFNNs with tanh activation function and $c_1 = 0.7$. The networks are both trained in the same LV but with different starting point: random policy for the orange line and BS optimal policy for the blue one. The dash-dotted horizontal red line is the conservative BS call option price computed according to the optimal risky portfolio in Figure 2 on the left. On the right agent actions in a never-seen MC episode.

Method	TVO price [EUR]
RL from random θ	$(1.912 \pm 0.003) \times 10^{-2}$
RL from BS θ	$(1.915 \pm 0.003) \times 10^{-2}$
Baseline	$(1.915 \pm 0.004) \times 10^{-2}$

Table 1: MC prices of an atm-spot TVO with LV dynamics for the risky assets and with $K = 1\,[\text{EUR}]$, $T = 1\,[\text{yr}]$ and $\bar{\sigma} = 5\%$. The prices are obtained computing three different allocation strategies: the ANN trained starting from a random policy, the ANN trained from the BS strategy and the baseline.

where the risky equities have different hedging costs, then the baseline and the BS solution gives the same TVO prices and optimal strategies; this is due to the fact that the numerator in Equation (28) plays an important role in the minimization and thus in the problem solution. For this reason we consider a market of two equities with two different shapes for the smiles curves and a μ_t vector that has equal positive entries. In particular we choose a symmetric smile for one equity and a displaced diffusion for the second one (Figure 5 on the left). In this way the BS optimal solution with long constraint depends only on the diffusive term:

$$\alpha^*(t) = \mu(t) \arg\min_{\alpha} \|\alpha \cdot \nu(t)\|^{-1}. \tag{56}$$

In a similar way the baseline

$$\alpha^*(t, S_t) = \mu(t) \arg\min_{\alpha} \|\alpha \cdot \nu(t, S_t)\|^{-1}. \tag{57}$$

Therefore, having a completely different smiles curves will make differ the BS solution from the

baseline one since the former is computed with a diffusive term $\nu(t)$ obtained by interpolating linearly the IV curve on the atm-spot strike and thus loosing all the information about the K-dependence. We train the RL agent in this market configuration by parameterizing the actions with the baseline strategy. In particular at each observational time $T_K \in \mathcal{T}$ the risky allocation strategy is obtained by summing the ANN output with the (57) and then remapped so that the long only strategy constraint is satisfied. We adopt this parameterization to help the agent find the optimal strategy as correction to the baseline one. We report in Figure 5 on the right the resulting learning curve compared with the TVO price adopting the state-independent BS strategy (56) and with the baseline (57). We observe that it learns as optimal policy the baseline strategy. We obtain the same result by adopting the second reward function (48) and decreasing the γ discount factor to make the agent prefer immediate rewards.

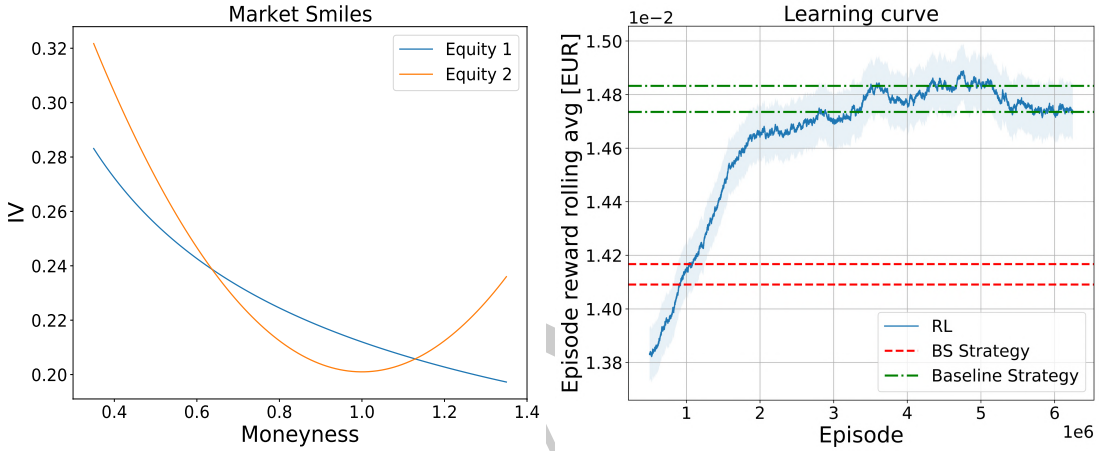

Figure 5: On the left market smiles for two equities in function of moneyness: symmetric smile at atm for equity 2 and displaced diffusion for equity 1. On the right learning curve for a 5×8 ANN trained in the LV market with the smile curves on the left. The learning curve is compared with the baseline and the static BS strategies.

6 Conclusion and Further Developments

In this paper we described a non-trivial control problem related to derivative contracts on target volatility strategies. We considered a bank 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, makes the bank solving a stochastic optimal control problem to adjust the price of the protection, since the bank strategy is not self-financing. We derived a closed form solution of the control problem in a Black and Scholes framework in two different ways: first by applying the Gyöngy Lemma and then by writing the Hamilton-Jacobi-Bellman equation. We studied the problem both in the BS world and in the local volatility one where the solution is not available and thus numerical investigation is needed. We tackled the problem by means of the novel Reinforcement Learning techniques, such as the proximal policy optimization. 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. Then we investigated the problem under LV dynamics. From our simulations we had evidence that under this model the RL agent learnt as optimal strategy the application path-wise of the BS solution. This topic is still under study since our results for the LV model do not provide a formal demonstration of that the problem solution is the path-wise BS one. For this reason the natural developments of this work is to compare our RL results with other standard techniques like the least-square Monte Carlo. If the resulting comparison is positive, then it will provide robustness of our LV results.

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.

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.

Derman, Emanuel and Iraj Kani (Jan. 1994). "Riding on a Smile". In: *Risk* 7.

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.

Dupire, Bruno (1994). "Pricing with a Smile". In: *Risk Magazine*, pp. 18–20.

Fleming, Wendell H. and Halil Mete Soner (2006). *Controlled Markov Processes and Viscosity Solutions*. Springer-Verlag New York.

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.

Gyöngy, I. (1986). "Mimicking the one-dimensional marginal distributions of processes having an ito differential". In: *Probability Theory and Related Fields* 71.4, pp. 501–516. DOI: 10.1007/bf00699039.

Halperin, Igor (2020). "QLBS: Q-Learner in the Black-Scholes(-Merton) Worlds". In: *The Journal of Derivatives* 28, pp. 99–122. ISSN: 1074-1240. DOI: 10.3905/jod.2020.1.108.

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.

Kolm, Petter N. and Gordon Ritter (2019). "Dynamic Replication and Hedging: A Reinforcement Learning Approach". In: *The Journal of Financial Data Science* 1.1, pp. 159–171. ISSN: 2640-3943. DOI: 10.3905/jfds.2019.1.1.159.

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.

Puheim, M. and L. Madarász (2014). "Normalization of inputs and outputs of neural network based robotic arm controller in role of inverse kinematic model". In: *2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pp. 85–89. DOI: 10.1109/SAMI.2014.6822439.

Schulman, John, Philipp Moritz, et al. (2016). "High-Dimensional Continuous Control Using Generalized Advantage Estimation". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun.

Schulman, John, Filip Wolski, et al. (2017). *Proximal Policy Optimization Algorithms*. arXiv: 1707.06347 [cs.LG].

Sola, J. and J. Sevilla (1997). "Importance of input data normalization for the application of neural networks to complex industrial problems". In: *IEEE Transactions on Nuclear Science* 44.3, pp. 1464–1468. DOI: 10.1109/23.589532.

Xue, Yuhong (Oct. 2012). *Target Volatility Fund: An Effective Risk Management Tool for VA?* Tech. rep. Society of Actuaries.